S&DS 230/530:
Data Exploration and Analysis

Ethan Meyers

Overview

Course overview
- Introductions
- syllabus/logistics

What the class covers

Review concepts from Intro Stats

Introduction to R
- R as a calculator
- Objects, vectors and data frames

Contact Information

Email: ethan.meyers@yale.edu
Office: 24 Hillhouse Ave, Room 206

Planned office hours: M 2:30-3:30pm, W 11-12

About me

Visiting assistant professor at Yale for this year
Assistant professor of Statistics Hampshire College
Research Fellow at the Center for Brains, Minds and Machines at MIT

Research: Machine learning to analyze neural data

Teaching Assistants

Teaching Fellows:

- Geyu Zhou: geyu.zhou@yale.edu
- Dylan O'Connell: dylan.oconnell@yale.edu
- Jiyi Liu: liu@yale.edu

Undergraduate Learning Assistants

- Coming soon…

Course objectives

1. To extend **methods** and **concepts** from intro stats to more complex real world settings

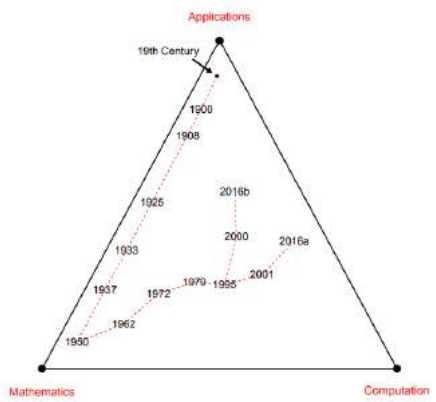
Gain insights on why/how particular methods work

- No math proofs, but we will explore concepts via computational simulations

2. To learn how to analyze and visualize **real data sets** using **the R programming language**

How to find the Truth/trends in a data set and convincingly convey the results to others!

How this class fits into S&DS



Efron and Hastie, Computer Age of Statistical Inference, 2016

Material based on previous versions of the class taught by Jonathan Reuning-Scherer, Susan Wang, Jay Emerson and Joseph Chang

Plan for the semester (subject to change)

First half: topics in traditional stats analyses

1	Aug 29	Course overview and introduction to R
2	Sep 3-5	Exploratory analysis and estimation with R
3	Sep 10-12	Confidence Intervals and the bootstrap
4	Sep 17-19	Hypothesis tests and permutation tests
5	Sep 24-26	Analysis of variance
6	Oct 1-3	Simple/multiple regression
7	Oct 8-10	Regression continued
8	Oct 15	Multi-factor analysis of variance
9	Oct 22-24	Midterm review and exam

Plan for the semester (subject to change)

Second half: Data Science approaches

10	Oct 28-30	Data wrangling
11	Nov 5-7	Data visualization
12	Nov 14-19	Mapping and joining data
13	Nov 19-22	Machine learning approaches
14	Dec 3-5	Misc topic, wrap up and review
15	Dec 13-18	Final exams

Topics we will cover

R and descriptive statistics/plots: The basics of base R, fundamental concepts in Statistics

Review confidence intervals: Sampling and bootstrap distributions, t-distributions

Review of hypothesis tests: Permutation tests, *non-parametric tests**, theories of testing

ANOVA: one-way/multi-way, interactions, *mixed effects**

Regression: simple/multiple, non-linear terms, logistic regression

Data wrangling: filtering and summarizing data, joining data sets, reshaping data

Data visualization: grammar of graphics, mapping

Statistical learning: cross-validation, *supervised learning*, PCA*, clustering*,*

***Misc:** text analysis and manipulation*, Bayesian methods*, other topics based on interest*

**Time permitting*

Examples of questions we might look at…

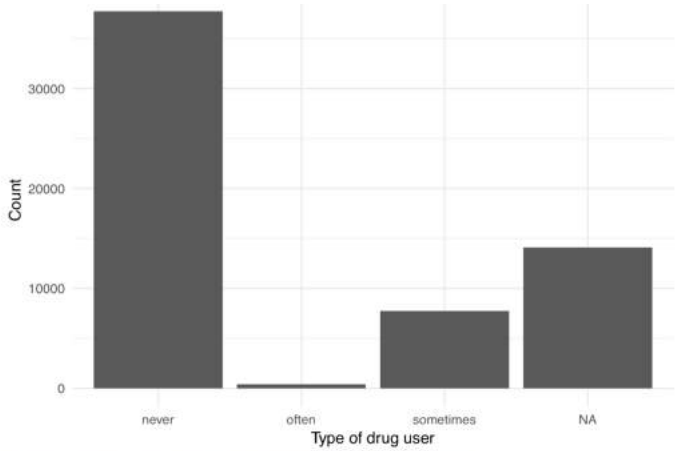
Randomization tests: Is it possible to smell whether someone has Parkinson's disease?

ANOVA: Are all genres of movies equally liked?

Data summarization: which airlines have the longest flight delays?

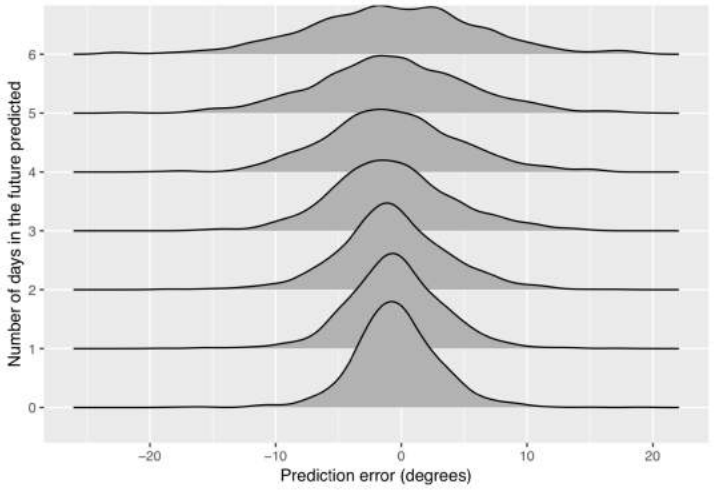
Data wrangling/visualization: How accurate are weather predictions?

Typical homework assignment piece




```
# Bonus: create a pie chart of the self reported frequency of
# drug use and make it look good!
profiles %>% count(drugs) %>% filter(!is.na(drugs)) %>% ggplot(aes(x = "",
    y = n, fill = drugs)) + geom_col(width = 1) + coord_polar(theta = "y") +
    theme_minimal() + theme(axis.title.x = element_blank(), axis.text.x = element_blank(),
    axis.ticks.x = element_blank(), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) + xlab("")
```

Typical homework assignment piece



Answers: Personally I like the joy plot best here because it most clearly shows how the distribution becomes more spread out for predictions made further in the future (although all three plots do a reasonable job of showing this).

Prerequisites

An introductory class in Statistics (AP or 10X)

- We will review Intro Stats concepts using computational methods but we will be going through the material at a fast pace

A large component of this class will be using the R programming

- No prior programming experience needed

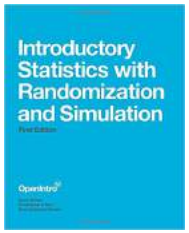
RStudio

DataCamp

Logistics

Website: https://yale.instructure.com/courses/51220

No required text, reading resources will be posted to canvas and in the homework assignments



Assignments and grades

1. Homework problem sets (60%)
   - Exploring concepts and analyzing data using R
   - Weekly: 10 total

Worksheet policies

- You may discuss questions with other but the work you turn in must be your own
- Worksheets assigned on Tuesdays and are due at 11:59pm on Sundays
- Late worksheets (90%) credit if turned in by 11:59pm on Monday
  - For any other extension a deans letter is needed
- Lowest scoring worksheet will be dropped

Assignments and grades

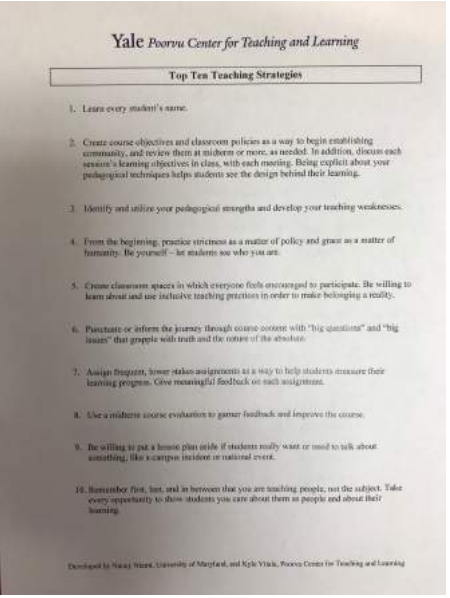
2. Examines (35%)
   - Midterm Oct 22nd or 24th (15%)
   - Final Dec 12-18th (20%) (or final project?)

3. Participation (5%)
   - Active asking and answering questions on Piazza

Class survey

In order for me to get to know you and to better adjust the class to your interests, please fill out the class survey on canvas
- Under the Quizzes link on the left

Any questions about the class logistics???



Yale *Poorvu Center for Teaching and Learning*

Top Ten Teaching Strategies

1. Learn every student's name.
2. Create course objectives and classroom policies as a way to begin establishing community, and review them at midterm or more, as needed. In addition, discuss each session's learning objectives in class, with each meeting. Being explicit about your pedagogical techniques helps students see the design behind their learning.
3. Identify and utilize your pedagogical strengths and develop your teaching weaknesses.
4. From the beginning, practice strictness as a matter of policy and grace as a matter of humanity. Be yourself – let students see who you are.
5. Create classroom spaces in which everyone feels encouraged to participate. Be willing to learn about and use inclusive teaching practices in order to make belonging a reality.
6. Punctuate or inform the journey through course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute.
7. Assign frequent, lower stakes assignments as a way to help students measure their learning progress. Give meaningful feedback on each assignment.
8. Use a midterm course evaluation to garner feedback and improve the course.
9. Be willing to put a lesson plan aside if students really want or need to talk about something, like a campus incident or national event.
10. Remember first, last, and in between that you are teaching people, not the subject. Take every opportunity to show students you care about them as people and about their learning.

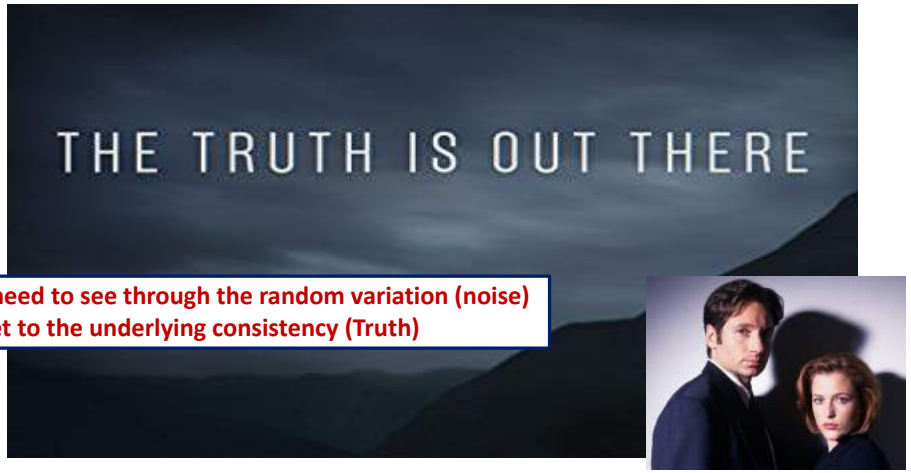
Developed by Nancy Vuont, University of Maryland, and Kyle Vitale, Poorvu Center for Teaching and Learning

Center for Teaching and Learning tips

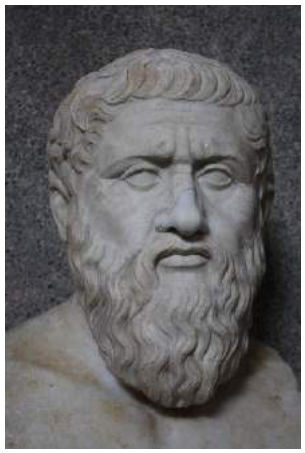
Tip 1: Learn every student's name

Tip 6: Punctuate or inform the journey through the course content with "big questions" and "big issues" that grapple with truth and the nature of the absolute

Quick Review of central concepts in Intro Statistics



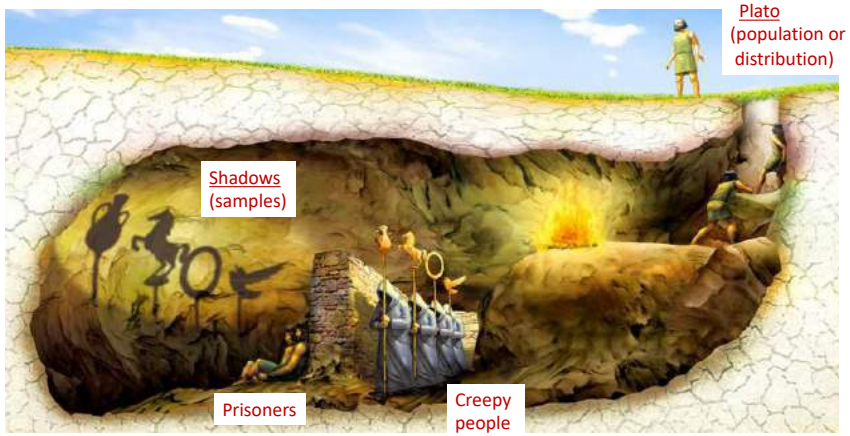
The Truth®!



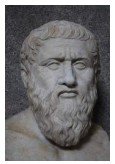
If we could see all the (infinite) data, we would know the Truth®!

Alas, we can only see a small subset of the data (a sample) so we merely see a shadow of the Truth

Plato's cave

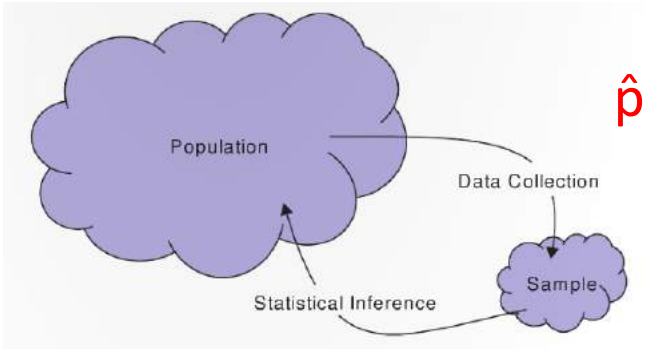


From The Republic (~ 380 BCE)



$\pi, \mu, \sigma, \rho, \beta$

Population: all individuals/objects of interest



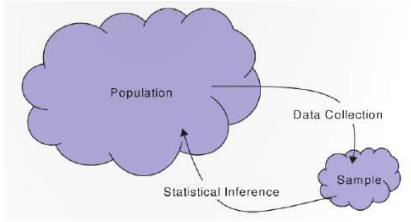
$\hat{p}, \bar{x}, s, r, b$

Sample: A subset of the population

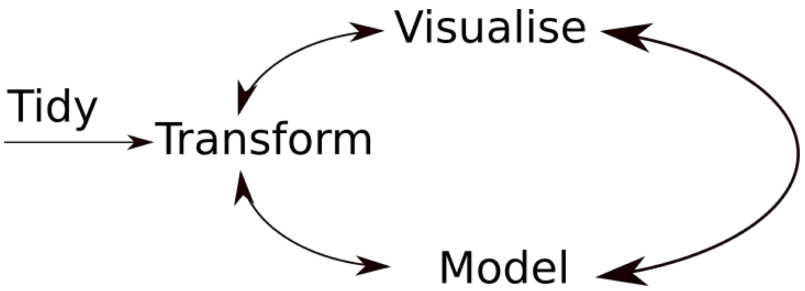
Descriptive and inferential statistics

Descriptive Statistics: describe the sample of data we have
- i.e., describe the shadows

Inferential Statistics: use the sample to make claims about properties of the population/process
- i.e., try to use the data to get at the Truth



Sometimes the truth is more complicated…



Question

Q: What programming language do pirates use?
A: Arrrr

Q: Worst joke of the semester?
A: Wait and see…

Log in to R Studio Cloud

bit.ly/SDS230_workspace_01

Or download and install R Studio

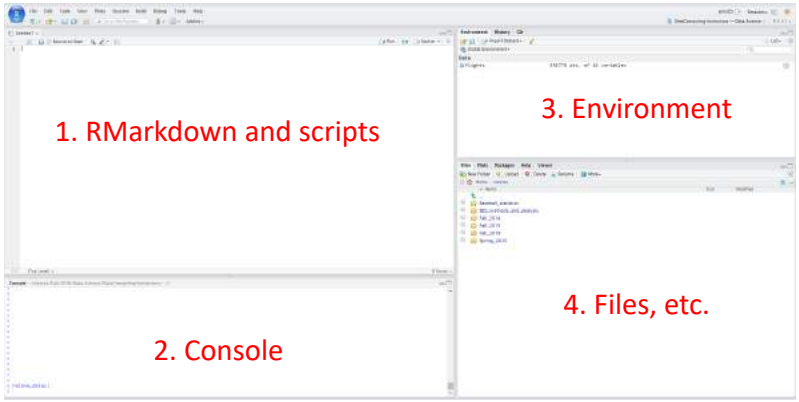
Talk to your neighbor while code is loading…

R and R Studio

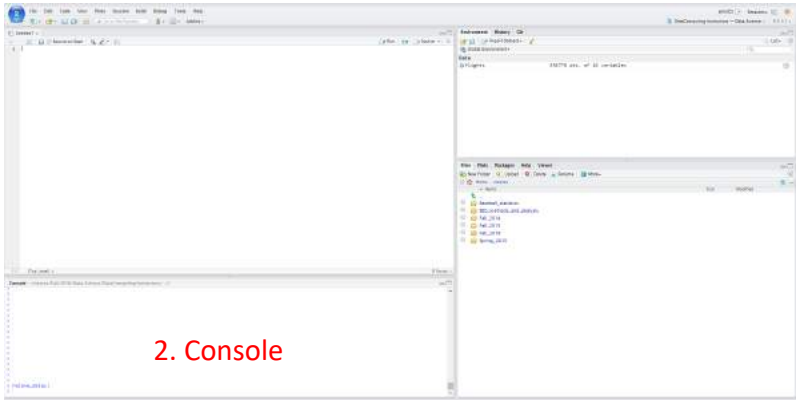
R: Engine

RStudio: Dashboard

RStudio layout



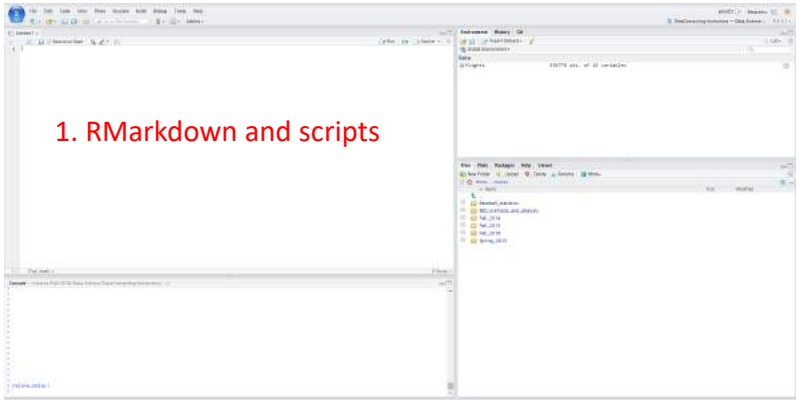
RStudio layout



R as a calculator

```
> 2 + 2
> 7 * 5
```

RStudio layout



Create a new script

File -> New File -> R Script

Save the script with a reasonable name, e.g., week1_notes.R

R Basics

Arithmetic:

```
> 2 + 2
> 7 * 5
```

Assignment:

```
> a <- 4
> b <- 7
> z <- a + b
> z
[1] 11
```

Number journey…

Number journey

```
> a <- 7
> b <- 52
> d <- a * b
> d
[1]  364
```

Character strings and booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE

> class(a)
[1] numeric

> class(s)
[1] character
```

Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
> tolower("DATA is AWESOME!")
```

To get help

```
> ? sqrt
```

One can add comments to your code

```
> sqrt(49)    # this takes the square root of 49
```

Vectors

Vectors are ordered sequences of numbers or letters
The c() function is used to create vectors

```
> v  <-  c(5, 232, 5, 543)
> s  <-  c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets []

```
> s[4]        # what will the answer be?
```

We can get multiple elements from a vector too

```
> s[c(1, 2)]
```

Vectors continued

One can assign a sequence of numbers to a vector

```
> z <- 2:10
> z[3]
```

One can test which elements are greater than a value

```
> z > 3
```

Can add names to vector elements

```
> names(v) <- c("first", "second", "third", "fourth")
```

Vectors continued

One can also apply functions to vectors

```
> z <- 2:10
> sqrt(z)
> mean(z)
```

Question

Q: What kind of grades the pirate get in Introduction to Statistics?
A: High Seas

Q: Worst joke of the semester?
A: Not likely

For next class

Fill out class survey on Canvas under the Quizzes link

Make sure R Studio cloud is working and/or set up R studio on your personal computer

Practice R and review Intro Stats concepts

- Resources page on Canvas: DataCamp, R videos

RMarkdown, data frames and plots

Overview

Very quick review
- Concepts
- R

R Markdown
- Formatting
- Code Chunks

More R
- Data frames
- Plots in base R

Logistics

R Studio Cloud outage happened during last class
- Current R Studio Cloud status
- Will keep trying for another week

Class background survey results

Logistics: homework 1

First homework problem set will be available after class

It is due on Sunday September 8th at 11:59pm
- 90% credit if turned by Monday September 9th at 11:59pm

The material on the problem set is based on what we will learn this week
- You should be able to answer questions 1, 3, and 5 by the end of today's class
  - Problem 5 is just reading and commenting on a blog post/Wired article

TA office hours are on a google calendar on Canvas

Logistics: homework 1

There are two ways to do the homework

1. The homework can be completed on R Studio Cloud (preferred)
https://rstudio.cloud/spaces/25704/project/477656

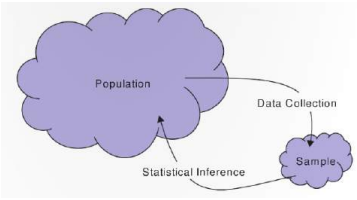
2. The homework should also work using R Studio desktop
   - Download the R Markdown document from Canvas
   - Run the code in the first chunk from the console

Warning: coding can be a bit frustrating at first
   - Try it on your own, then get help as needed (Piazza, TA office hours)

Very quick review

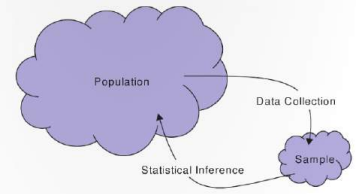
Concepts from Intro Stats

The Truth® is out there
   - If we have infinite data we could compute parameters
   - We can estimate parameters with statistics
      - statistics are functions of our data



	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
regression slope	b	β

Sampling



Simple random sample: each member in the population is equally likely to be in the sample
   - This is called ***random selection***

Soup analogy!

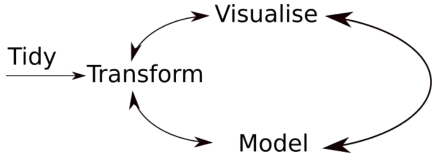
Q: Why is this good?

A: Allows for generalizations to the population! (no bias)

Our goal

Our goal is to try to uncover the Truth and convey the results to others
   - More than one way to get at the Truth
      - no one 'right way' to analyze data
   - Many ways to make mistakes too

Getting at the Truth can be an iterative process



"All models are wrong
some models are useful"
- George Box

Review of R from last class

Assignment:

```
> a <- 5
```

Data types:

```
> s <- "s is a terrible name for an object"    # string
> b <- TRUE     # boolean
```

Functions:

```
> sqrt(49)
> ? sqrt    # help
```

Review of R from last class

Vectors:

```
>  v  <-  c(TRUE, TRUE, FALSE)        # a vector of booleans
```

Accessing elements of a vector:

```
>  v[3]
>  v[c(2, 3)]
```

RMarkdown

RMarkdown (.Rmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

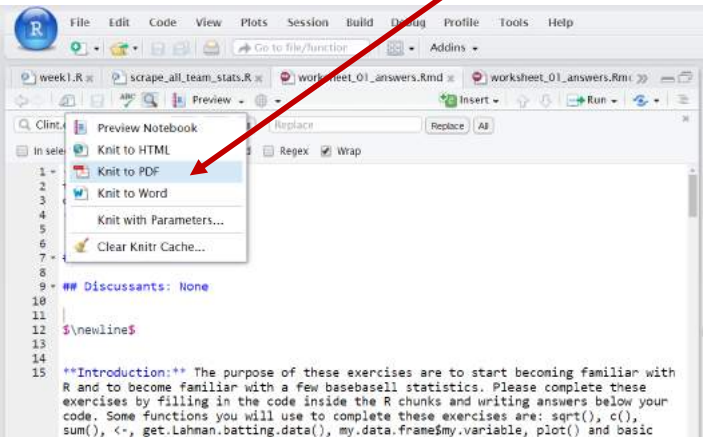
Creates a way to do reproducible research!

Boot up R Studio to follow along:

- Either on your own computer or on R Studio Cloud:
  https://rstudio.cloud/spaces/25704/project/481362

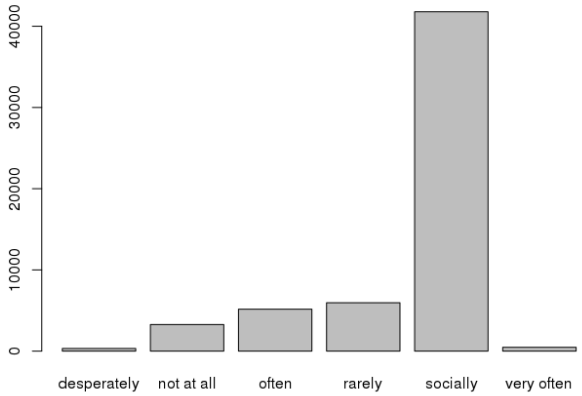
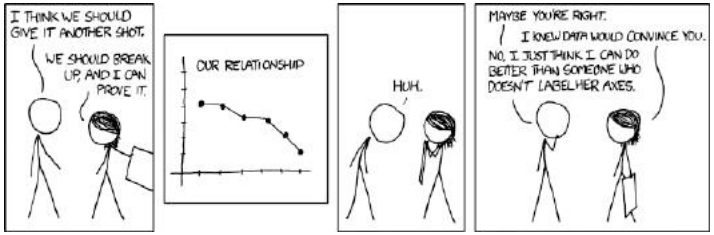
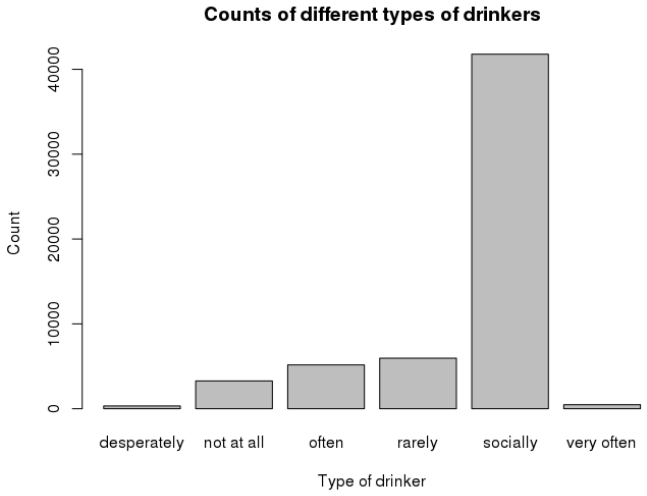
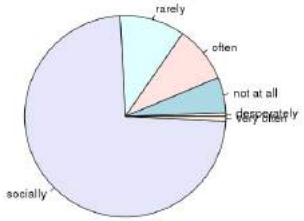
RMarkdown

Everything in R chunks is executed as code:

```
```{r}
    # this is a comment
    # the following code will be executed
    2 + 3
```
```

Everything outside R chunks appears as text

Knitting to a pdf

Turn in a pdf or html document with your solutions to Canvas



RMarkdown

Note: When you knit, RMarkdown files **do not have access to variables in the global environment**, but instead have their own environment.

Why is this a good thing???

Formatting in R Markdown

We can add formatting to text outside the code chunks

Examples:

```
## Level 2 header
**bold**
![](https://statistics.yale.edu/sites/default/files/logo2.png)
```

LaTeX in R Markdown

We can also add LaTeX symbols to documents using $\symbol$ syntax

For example, try these:

```
$\theta$
$\hat{p}$
$\hat{\theta}$
```

Knit early and knit often to avoid errors!!!

LaTeX in R Markdown

I have added a link on Canvas in the resources section to help find LaTeX symbols

How else could you get help to learn more about LaTeX symbols?

To repeat: avoid hard to debug code!

Only change a few lines at a time and then knit your document to make sure everything is working!

If you document isn't knitting:

- **For code:** use the # symbol until you can find the line of code that is giving the error message
- **For syntax:** cut part of the document until it knits and then paste it back

Back to R: Data frames

Data frames contain structured data

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	*NA*	graduated from masters program
4	23	thin	vegetarian	socially	*NA*	working on college/university
5	29	athletic	*NA*	socially	never	graduated from college/university
6	29	average	mostly anything	socially	*NA*	graduated from college/university

OK Cupid data

Back to R: Data frames

Data frames contain structured data

```
# install.packages("okcupiddata")    # only needs to be run once
> library(okcupiddata)
> View(profiles)      # the View() function only works in R Studio!
```

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	*NA*	graduated from masters program
4	23	thin	vegetarian	socially	*NA*	working on college/university
5	29	athletic	*NA*	socially	never	graduated from college/university
6	29	average	mostly anything	socially	*NA*	graduated from college/university

Data Frames

Variables

Cases

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	*NA*	graduated from masters program
4	23	thin	vegetarian	socially	*NA*	working on college/university
5	29	athletic	*NA*	socially	never	graduated from college/university
6	29	average	mostly anything	socially	*NA*	graduated from college/university

An Example Dataset (Shadows)

Quantitative Variable

Categorical Variable

Cases (observational units)

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	*NA*	graduated from masters program
4	23	thin	vegetarian	socially	*NA*	working on college/university
5	29	athletic	*NA*	socially	never	graduated from college/university
6	29	average	mostly anything	socially	*NA*	graduated from college/university

Data frames

When data is loaded from a package it isn't visible in the environment pane. We can make it visible using the data() function.

```
> library(okcupiddata)
> data(profiles)
```

We can extract the columns of a data frame as vector objects using the $ symbol

```
> the_ages <- profiles$age
```

Data frames

```
> the_ages <- profiles$age
```

Can you get the mean() age of users in this data set?

```
> mean(the_ages)
```

Extracting rows from a data frame

We can extract rows from a data frame in a similar way as extracting values from a vector by using the square brackets

```
> profiles[1, ]  # returns the first row of the data frame
> profiles[, 1]  # returns the first column of the data
```

Note, the first column of the profiles data frame is the variable *age*, so we can also get the first column using:

```
> profiles$age  # this is the same as profiles[, 1]
```

Extracting rows from a data frame

We can also create vectors of numbers or booleans specifying which rows we want to extract from a data frame

```
# create a vector with the numbers 1, 10, 20
> my_vec <- c(1, 10, 20)

# use my_vec to get the 1st, 10th, and 20th row in profiles
> small_profiles <- profiles[my_vec, ]
> dim(small_profiles)  # number of rows and columns in the data frame
```

Extracting rows from a data frame

Finally, we can also extract rows by creating a Boolean vector that is of the same length as the number of rows in the data frame

TRUE values will be extracted from the data frame while FALSE values will not

```
# create a vector of booleans
> my_bools <- c(TRUE, FALSE, TRUE)

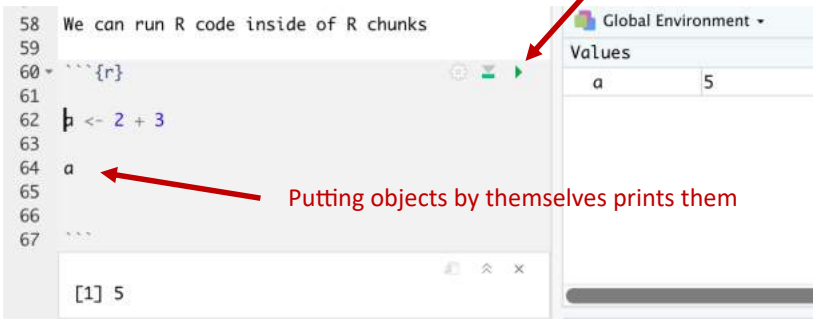
# use the Boolean vector to get the 1st and 3rd row
> my_bools <- c(TRUE, FALSE, TRUE)
> small_profiles[my_bools, ]
```

Categorical data

Categorical variables take on one of a fixed number of possible values

For categorical variables we usually want to view:

- How many items are each category or
- The proportion (or percentage) of items in each category

```
# Get information about drinking behavior
> drinking_vec <- profiles$drinks

# Create a table showing how often people drink
> drinks_table <- table(drinking_vec)
> drinks_table
```

Relative frequency table

We can create a relative frequency table using the function:

```
> prop.table(my_table)
```

Can you create a relative frequency table for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
> prop.table(drinks_table)
```

Bar plots (no pun intended)

We can plot the number of items in each category using a bar plot

```
> barplot(my_table)
```

Can you create a bar plot for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
> barplot(drinks_table)
```



What is wrong with this plot?

- A: the axes are not labeled!!!



If you don't want exes, label you axes!

Details matter!

Can you figure out how to label the axes?

- A: ? barplot
- A: xlab and ylab!

```
> barplot(drinks_table,
        ylab = "Count",
        xlab = "Type of drinker",
        main = "Counts of different types of drinkers")
```



So much better!!!

Pie charts

We can also use the pie() function to create pie charts

> pie(drinks_table)



Which is best bar plots or pie charts?

> barplot(table(profiles$sex, useNA = "always"))

> pie(table(profiles$sex, useNA = "always"))

Q1: Is one better than the other?

Q2: Can you figure out how to add colors to these plots?

Homework 1

Homework 1 due on Sunday September 8th at 11:59pm

TA office hours are on a google calendar on Canvas

Plots continued, for loops, sampling distributions

Overview

Very quick review
- R from last class

Continue with more R
- Plots of categorical data continued
- Plots of quantitative data
- For loops
- Sampling distributions

Announcements

We now have 4 undergraduate learning assistants for this class!

- Dalton Boyt: dalton.boyt@yale.edu
- Derek Chen: derek.chen@yale.edu
- Jessica Pevner: jessica.pevner@yale.edu
- Maria Eduarda Santana: mariaeduarda.santana@yale.edu

Their office ours are on the calendar on Canvas

Dylan's office hours: 1-3PM – is Tuesday or Thursday better?
I will have extra office hours today from 1:30-2:30

Logistics: homework 1

Reminder homework 1 is due Sunday September 8th at 11:59pm

Has anyone started on it yet?
- How is it going?

Glad to see people are using Piazza
- Intro Stats: The Truth is out there (statistics are estimate of parameters)
- This class: The truth is more complicated

Review of R from last class

R Markdown syntax

```
**bold**
$\LaTeX$
```

Code chunks:

```
```{r}
  2 + 3
```
```

knit often!

Review of R from last class

Use R Markdown interactively



- When using the green play button, R Markdown has access to environment objects
- Chunks have access to objects created in earlier chunks

Review of R from last class

Data frames:

> View(profiles) # doesn't work in R Markdown documents

> the_ages <- profiles$age

> mean(the_ages)

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review of R from last class

Extracting subsets of data from data frames:

> profiles[1,] # returns the first row of the data frame

> small_profiles <- profiles[c(1, 10, 20),]

> small_profiles[c(TRUE, FALSE, TRUE),]

Creating tables for categorical data

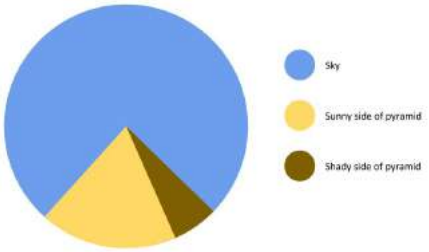
> drinks_table <- table(drinking_vec)

> prop.table(drinks_table)

Review of R from last class

Plotting categorical data

```
> barplot(drinks_table,
        ylab = "Count",
        xlab = "Type of drinker",
        main = "Counts of drinkers")
```



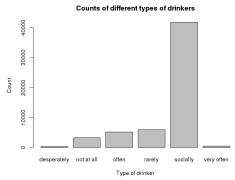
> pie(drinks_table)

Which is best bar plots or pie charts?

```
> barplot(table(profiles$sex, useNA = "always"))
> pie(table(profiles$sex, useNA = "always"))
```

Q: Can you figure out how to add colors to these plots?

Removing social drinkers



Social drinkers are dominating our plot ☹

We can get rid of social drinkers by only plotting counts less than 10,000

```
> nonsocial_inds <- drinks_table < 10000
> nonsocial_drinks_table <- drinks_table[nonsocial_inds]
> barplot(nonsocial_drinks_table)
```

Quantitative data: statistics

What are some statistics that describe the central tendency of quantitative data?

- The mean $\bar{x}$: mean()
- The median m: median()

Which of these measures is robust to outliers?

Can you calculate the mean and median of OkCupid user's heights?

Quantitative data: Visualizing heights

Q: How can we visualize the heights in the profiles data frame?

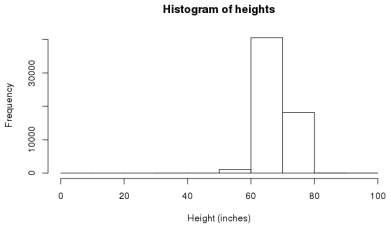
A: Histograms!

A: Boxplots

A: Many other options too

Histograms of heights

Height (inches)	Frequency Count
(0-10]	6
(10-20]	0
(20-30]	1
(30-40]	13
(40-50]	9
(50-60]	1097
(60-70]	40575
(70-80]	18164
(80-90]	50
>90	28



Visualizing heights

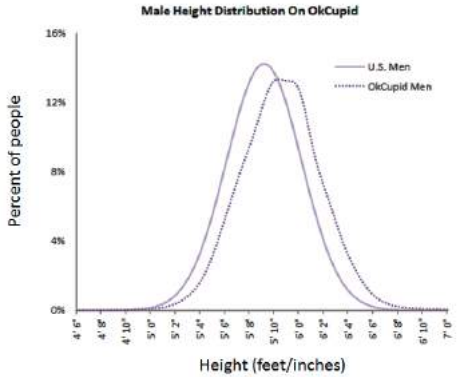
We can create histograms in R using the hist() function

Can you create a histogram of heights?

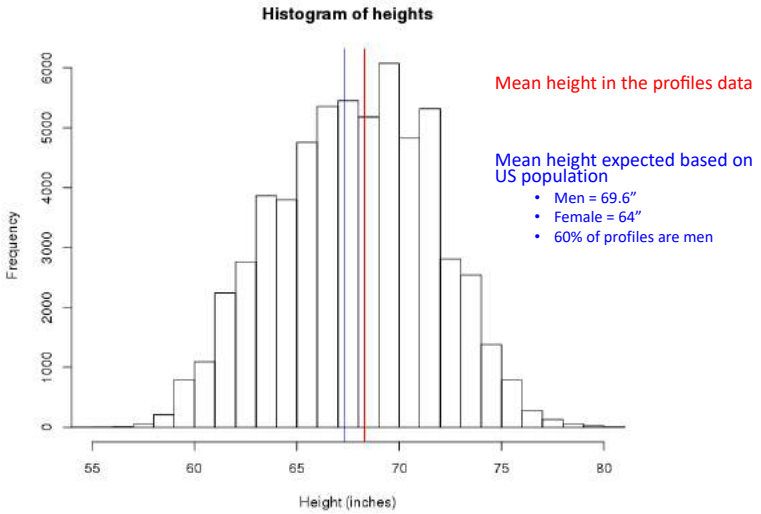
> hist(profiles$height)

> hist(profiles$height, nclass = 50)

OkCupid users are taller than the average person



Can we see this in the profiles data?



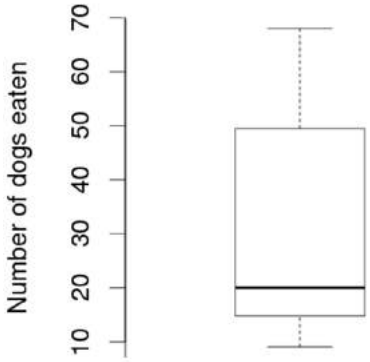
Mean height in the profiles data

Mean height expected based on US population

- Men = 69.6”
- Female = 64”
- 60% of profiles are men

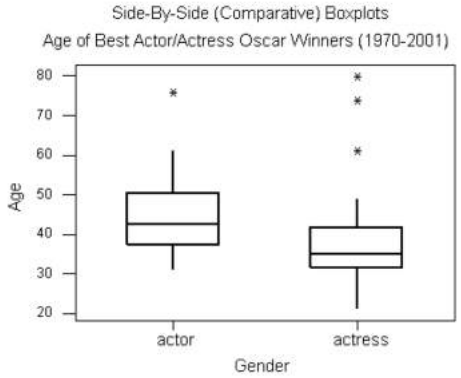
abline() adds lines to plots

Box plots can also visualize quantitative data



R: `boxplot(v)`

Side-by-side boxplots



Useful for comparing distributions!

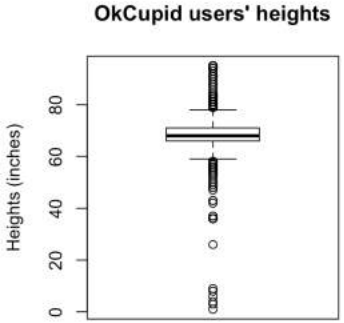
- What does the figure above show?

R: `boxplot(v1, v2)`

Outliers

Outliers on boxplots are values that are more than 1.5 * IQR

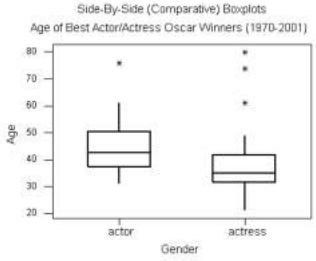
What should we do if we have outliers?



Outliers

Outliers on boxplots are values that are more than 1.5 * IQR

What should we do if we have outliers?



CitiBike data

Let's look at the bike share data from NYC

```
> load('daily_bike_totals.rda')
```

CitiBike data

Let's look at the bike share data from NYC

```
> load('daily_bike_totals.rda')
```

CitiBike analysis

What does each case correspond to?

We can use the dim() function to get how many cases and variables there are

- How many are there?

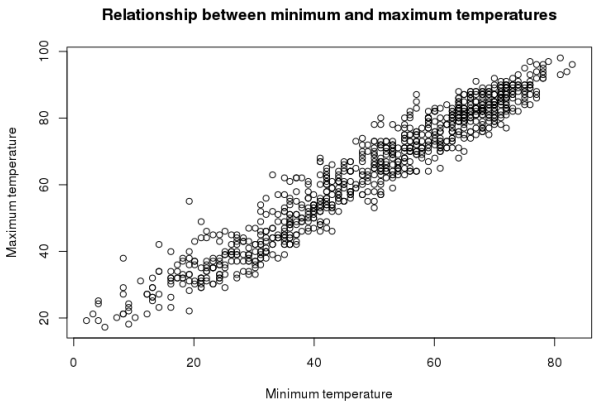
Scatter plots

We can use the plot(x, y) function to create scatter plots

Can you create a scatter plot of the relationship between the minimum and maximum temperatures?

```
> plot(bike_daily_data$min_temperature,
       bike_daily_data$max_temperature,
       xlab = "Minimum temperature",
       ylab = "Maximum temperature",
       main = "Relationship between min and temp")
```

Scatter plots



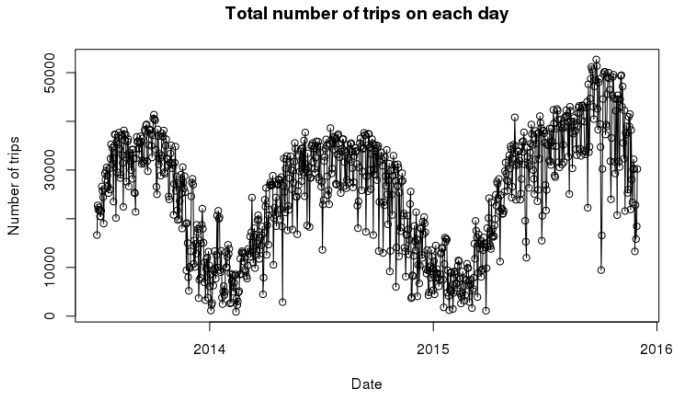
Plotting time series

We can use the plot(x, y) function to plot time series

```
# we can connect the points in a plot using
> plot(x, y, type = 'l')   # connected points
> plot(x, y, type = 'o')   # both points and dots

> plot(bike_daily_data$date,  bike_daily_data$trips,
        type = 'o',
        xlab = "Date",
        ylab = "Number of trips",
        main = "Total number of trips on each day")
```

Plotting time series



For loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

The syntax for a for loop is:

```
for (i in 1:100) {
        # do something
}
```

← This is repeated 100 times
i is incremented by 1 each time

For loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

The syntax for a for loop is:

```
for (i in 1:100) {
        print(i)
}
```

← This is repeated 100 times
i is incremented by 1 each time

For loops

For loops are particular useful in combination with vectors that can store the results

```
my_results <- NULL      # create an empty vector to store the results
for (i in 1:100) {
        my_results[i] <- i^2
}
```

Sometimes there are more efficient ways to do the same thing without for loops

```
> (1:100)^2
```

Generating random data

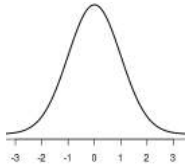
R has built in functions to generate data from different distributions

The uniform distribution:

```
# generate n = 100 points from U(0, 1)
        rand_data <- runif(100)
        hist(rand_data)
```

The normal distribution

```
# generate n = 100 points from N(0, 1)
        rand_data  <- rnorm(100)
        hist(rand_data)
```



Generating random data

If we want the same sequence of random numbers we can set the random number generating seed

```
> set.seed(123)
> runif(100)
```

Q: Why would we want the same sequence of random number?

Sample statistics

Q: What is a statistic?

The sample mean $\bar{x}$ (shadow of the parameter μ)

```
rand_data <- runif(100)    # generate n = 100 points from U(0, 1)
mean(rand_data)
```

Q: If we repeat the code above will we get the same statistic?

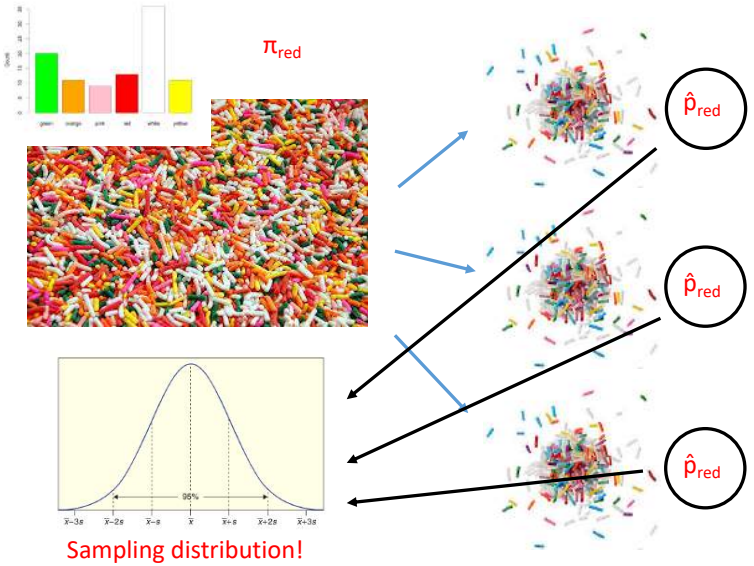
Sampling distribution

A distribution of ***statistics*** is called a ***sampling distribution***

Reminder: For a *single* ***categorical variable***, the main statistic of interest is the ***proportion*** ($\hat{p}$) in each category

- (shadow of the parameter π)

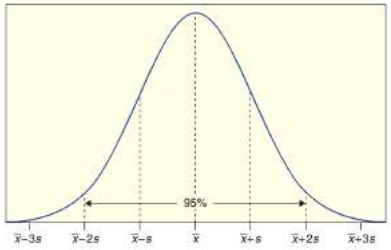
$\hat{p}$ = Proportion in a category = $\frac{\text{number in that category}}{\text{total number}}$



Sampling distribution

Q1: Would we ever calculate the sampling distribution in practice?

Q2: Why would we be interested in the sampling distribution?



Sampling distribution

Sampling distributions

```
sampling_dist <- NULL
for (i in 1:1000) {
        rand_data <- runif(100)    # generate n = 100 points from U(0, 1)
         sampling_dist[i] <- mean(rand_data)    # save the mean
}

hist(sampling_dist)
```

Sampling distributions

Distribution of OkCupid user's heights n = 100

```
heights <- profiles$height

# get one random sample of heights from 100 people
height_sample <- sample(heights, 100)

# get the mean of this sample
mean(height_sample)
```

Sampling distributions

Distribution of OkCupid user's heights n = 100

```
sampling_dist <- NULL
for (i in 1:1000) {
        height_sample <- sample(heights, 100)    # sample 100 random heights
        sampling_dist[i] <- mean(height_sample)    # save the mean
}

hist(sampling_dist)
```

Homework 1

Homework 1 due on Sunday September 8th at 11:59pm

TA office hours are on a google calendar on Canvas

Confidence intervals and the bootstrap

Point Estimate

We use the statistics from a sample as a **point estimate** for a population parameter

- $\bar{x}$ is a point estimate for…?

44% of American approve of Trump's job performance

- Gallup poll from October 14th

Symbols:

π: Trump's approval for all voters

$\hat{p}$: Trump's approval for those voters in our sample

Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a population parameter

One common form of an interval estimate is:

Point estimate ± margin of error

Where the **margin of error** is a number that reflects the precision of the sample statistic as a point estimate for this parameter

Example: Fox news poll

44% of American approve of Trump's job performance, plus or minus 3%

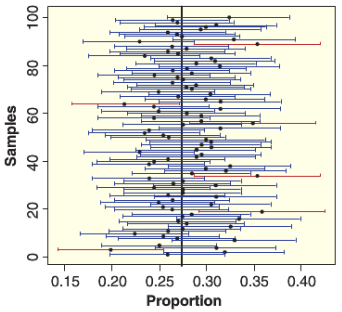
How do we interpret this?

Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the ***parameter*** a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter



Think ring toss…

Parameter exists in the ideal world

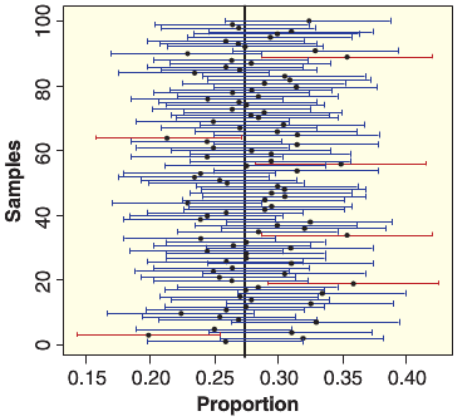
We toss intervals at it

95% of those intervals capture the parameter

Confidence Intervals

For a **confidence level** of 95%...

95% of the **confidence intervals** will have the parameter in them



Wits and Wagers…

Wits and Wagers…

Question 1: In feet and inches, how tall was the tallest human in recorded history?

Question 2: How many floors does the leaning tower of Pisa have?

Question 3: What year was the parking meter invented?

Wits and Wagers…

Question 4: How many time zones does Russia have?

Question 5: What percentage of US households own a cat?

Question 6: What percent of the world's population lives in the U.S.?

Question 7: On average, what percent of a watermelon's weight comes from water?

Wits and Wagers…

Question 8: How many chemical elements are there on the Periodic Table of the Elements?

Question 9: What percent of the world's surface is water?

Question 10: In what year was an ATM machine first installed in the U.S.?

Note

For any given confidence interval we compute, we don't know whether it has really captured the parameter

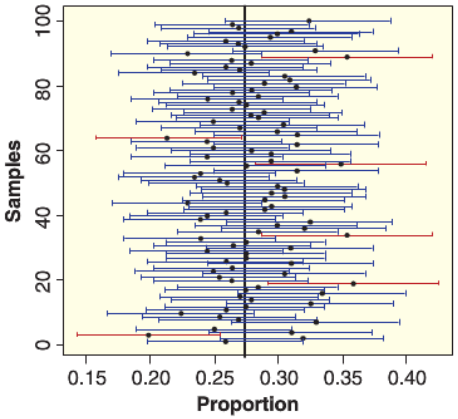
But we do know that if we do this 100 times, 95 of these intervals will have the parameter in it

(for a 95% confidence interval)

Confidence Intervals

For a **confidence level** of 90%...

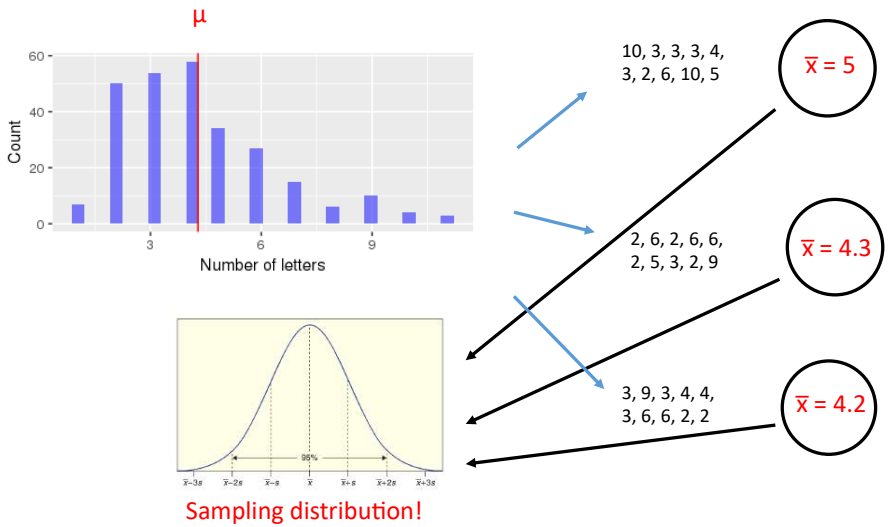
90% of the **confidence intervals** will have the parameter in them



Computing confidence intervals

Let's now discuss how we can compute confidence intervals…

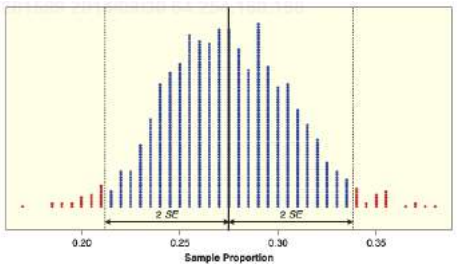
Review: sampling distribution illustration



Sampling distribution!

Sampling distributions

Q: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?

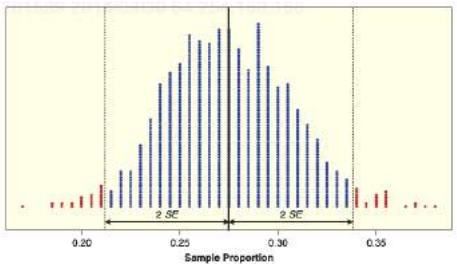


Sampling distributions

Q: If we had:

- A statistics value
- The SE

Could we compute a 95% confidence interval?



Sampling distributions

Q: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

Sampling distributions

Q: If we can't calculate the sampling distribution, what's else could we do?

Plug-in principle

Suppose we get a sample from a population of size *n*

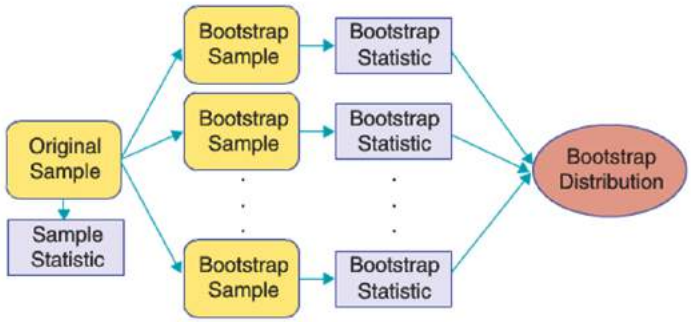
We pretend that *the sample is the population* (plug-in principle)

1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

2. We repeat this process 1000's of times and get a ***bootstrap sample distribution***

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Bootstrap process



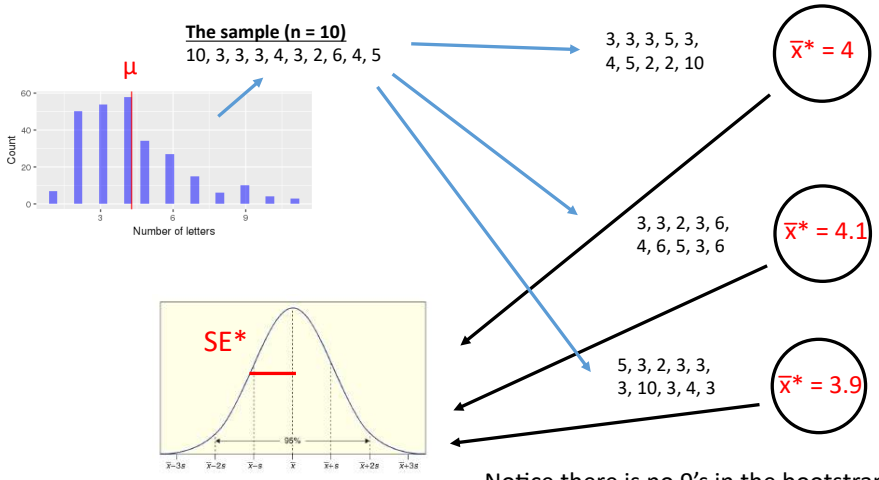
95% Confidence Intervals

When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

Bootstrap distribution illustration



Let's try it in R...

Confidence intervals and the bootstrap

Confidence Intervals

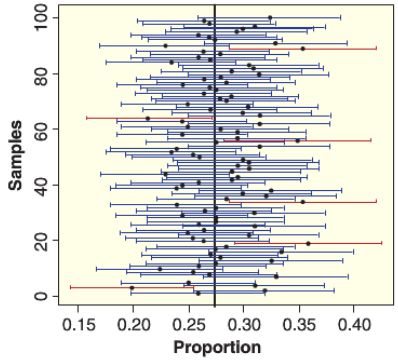
A **confidence interval** is an interval computed by a method that will contain the ***parameter*** a specified percent of times

The **confidence level** is the percent of all intervals that contain the parameter

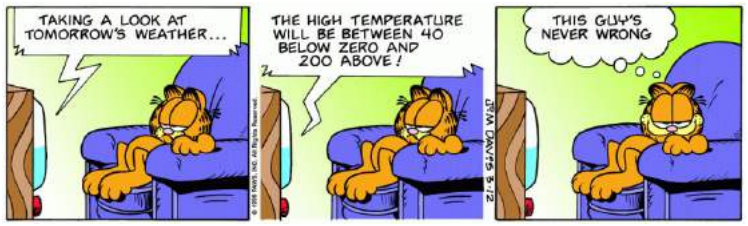
Confidence Intervals

Q: For a **confidence level** of 90%, how many of these intervals should have the parameter in them?

Q: For a given confidence interval, do we know if it contains the parameter?



Q: For the cartoon below, what is the confidence level the weatherman is using?



There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Example

130 observations of body temperature of men were made (°F)

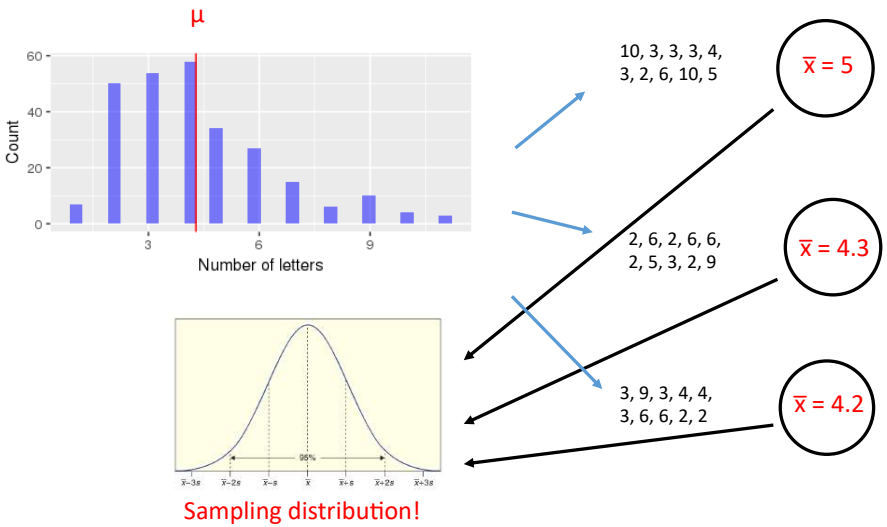
A 95% confidence interval for the mean body temperatures is:
[98.123, 98.375]

How do we interpret these results?

Is this what you would expect?

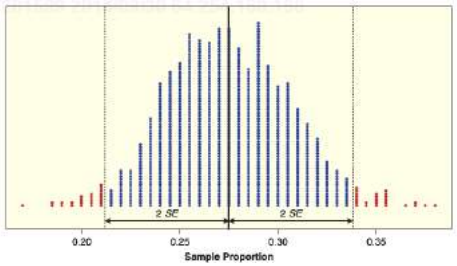
Original paper Statistic teaching resource

Review: sampling distribution illustration



Sampling distributions

Q: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?



If we had:

- A statistics value
- The SE

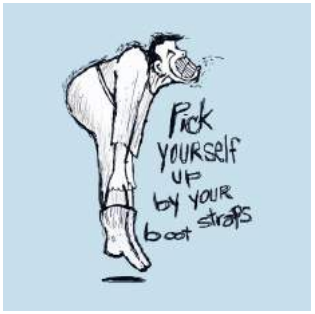
We could compute a 95% confidence interval!

Sampling distributions

Unfortunately we can't calculate the sampling distribution ☹

We have to pick ourselves up by the bootstraps!

1. Estimate SE with $\hat{SE}$
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI



Plug-in principle

Suppose we get a sample from a population of size *n*

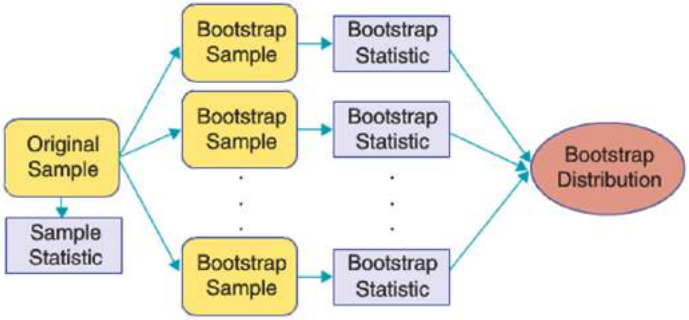
We pretend that *the sample is the population* (plug-in principle)

1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

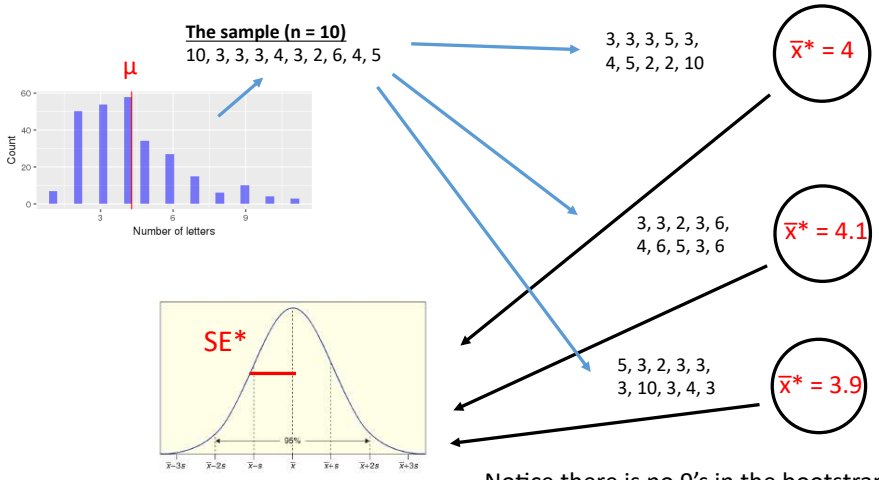
2. We repeat this process 1000's of times and get a ***bootstrap sample distribution***

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Bootstrap process



Bootstrap distribution illustration



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

95% Confidence Intervals

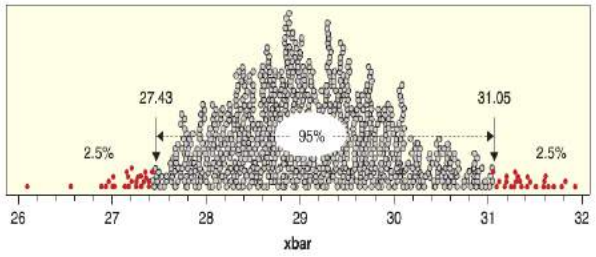
When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

What if the bootstrap distribution is not normal?

If the bootstrap distribution is approximately symmetric, we can use percentiles in the bootstrap distribution to an interval that matches the desired confidence level.



Findings CIs for many different parameters

This bootstrap method works for constructing confidence intervals for many different types of parameters!

Let's try it in R…

Formulas for the standard error of the mean

As you likely learned in intro statistics class, there is formula the **standard error of the mean** which is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where:

- σ is population standard deviation parameter
- n is the sample size
- s is the sample standard deviation

Formula for the standard error of a proportion

Likewise, there is a formula for **standard error of a proportion** which is:

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \qquad s_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Where:

- $\hat{\pi}$ is the population proportion parameter
- n is the sample size
- $\hat{p}$ is the sample proportion statistic

Next class, hypothesis tests…

Review of hypothesis tests

Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population

Example 1: we might make the claim that Trump's approval rating for all US citizens is 45%

How can we write this using symbols?

Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population

Example 2: we might make the claim that the average height of a baseball player is 72 inches

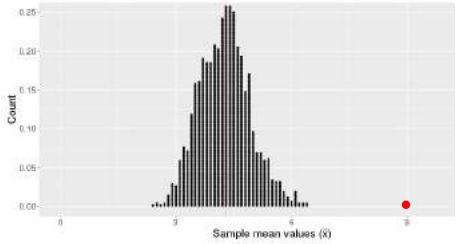
How can we write this using symbols?

Basic hypothesis test logic

We start with a claim about a population parameter

- E.g., $\mu = 4.2$

This claim implies we should get a certain distribution of statistics



If our observed statistic is highly unlikely, we reject the claim

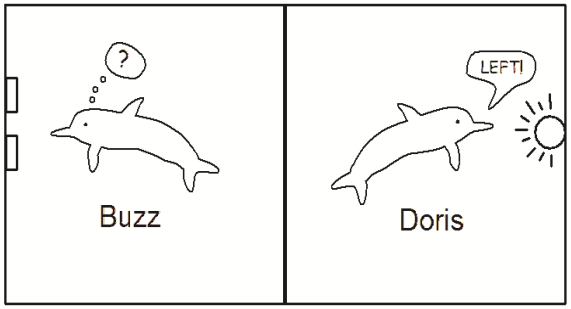
Are dolphins capable of abstract communication?

Dr. Jarvis Bastian is the 1960's wanted to know whether dolphins are capable of abstract communication

He used an old headlight to communicate with two dolphins (Doris and Buzz)

- Stead light = push button on right to get food
- Flashing = push button on the left to get food

A canvas was then put in the middle of the pool with Doris on one side and buzz on the other



Left

Right

Buzz got 15 out of 16 trials correct

Questions about the experiment

1. What are the cases here?

2. What is the variable of interest and is it categorical or quantitative?

3. What is the observed statistic - and what symbols should we use to denote it?

4. What is the population parameter we are trying to estimate - and what symbol should we use to denote it?

5. Do you think the results are due to chance?
   - i.e., how many correct answers do you think Buzz would have gotten if he was guessing?

6. Are dolphins capable of abstract communication?

The dolphin communication study

If Buzz was just guessing, what would we expect the value of the parameter to be?

If Buzz was not guessing, what would we expect the value of the parameter to be?

Chance models

How can we assess whether 15 out of 16 correct trials ($\hat{p}$ = .975) is beyond what we would expect by chance?

Chance models

To really be sure, how many repetitions of flipping a coin 16 times should we do?

Any ideas how to do this?

How to simulate coin flips in R

We can simulate coin flipping using the rbinom() function

flip_simulations <- rbinom(num_sims, size, prob)

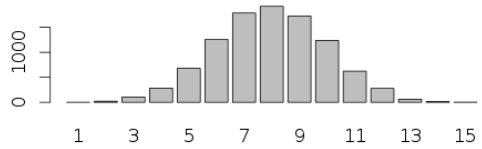
num_sims: the number of simulations run
- Typically we do around 10,000 repeats

size: the number of trials on each simulation
- 16 for Doris/Buzz

prob: the probability of success on each trial
- .5 Doris/Buzz

Simulating Flipping 16 coins 10,000

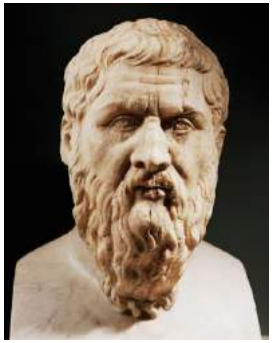
0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0



Is it likely that Buzz was guessing?

Are dolphins capable of abstract communication?

Question: who is this?



Question: who is this?

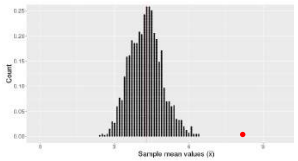
To prove G wrong, we will start by assuming he is right!

Namely, we will assume H_0 (that $\pi = 0.5$)

We will then generate a number of statistics ($\hat{p}$) that are consistent with H_0

- i.e., we will create a ***null distribution***

If our observed statistic looks very different from the statistics generated under we can reject H_0 and accept H_A

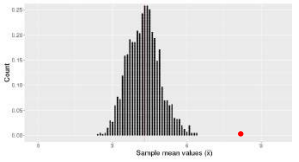


Terminology

Null Hypothesis (H_0): Claim that there is no effect or no difference

Alternative Hypothesis (H_A): Claim for which we seek significant evidence

The alternative hypothesis is established by observing evidence that inconsistent with the null hypothesis



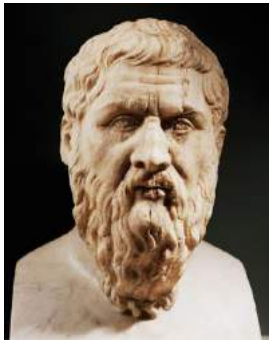
Review: the dolphin communication study

1. What is the null hypothesis?

2. We can write this in terms of the population parameter as:

3. What is the alternative hypothesis?

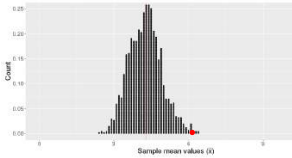
Setting the rules



Life wisdom: If you are going to make a bet with a a nihilist, you'd better agree to the rules first!

Rules

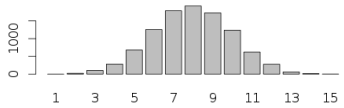
- If there is a less than 5% chance we would get a random statistic as or more extreme than the observed statistic (if H_0 is true) we will reject H_0
  - i.e., Gorgias loses the bet
- In symbols: $\alpha = 0.05$



Null Distribution

A **null distribution** is the distribution of statistics one would expect if the null hypothesis (H_0) was true

i.e., the null distribution is the statistics one would expect to get if nothing interesting was happening



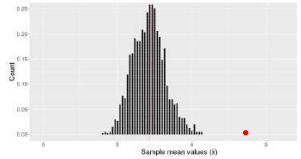
P-values

A **p-value** is the probability, of obtaining a statistic as as (or more) extreme than the observed sample *if the null hypothesis was true*

- i.e., the probability that we would get a statistic as extreme as our observed statistic from the null distribution

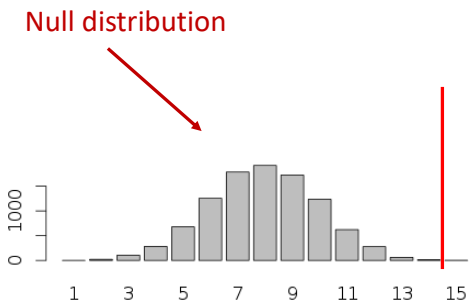
$$\Pr(\text{STAT} \geq \text{observed statistic} \mid H_0 = \text{True})$$

The smaller the p-value, the stronger the statistic evidence is against the null hypothesis



Buzz and Doris example

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0



p-value = 3/10000 = .0003

Statistical significance

When our observed sample statistic is unlikely to come from the null distribution, people often say the results are **statistically significant**

- i.e., our p-value is less than α

'Statistically significant' results mean we have strong evidence against H_0 in favor of H_a

- The American Statistical Association rejects the phrase 'statistically significant'

Key steps hypothesis testing

1. State the null hypothesis… and the alternative hypothesis
2. Calculate the observed statistic
3. Create a null distribution that is consistent with the null hypothesis
4. Examine how likely the observed statistic is to come from the null distribution
5. Make a judgement

Is it possible to smell whether someone has Parkinson's disease?

Joy Milne claimed to have the ability to smell whether someone had Parkinson's disease

To test this claim researchers gave Joy 6 shirts that had been worn by people who had Parkinson's disease and 6 people who did not

Joy identified 11 out of the 12 shirts correctly

Question: Can Joy really smell whether someone has Parkinson's disease?

Let's examine this in R…

Review of hypothesis tests

Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population

Example 1: we might make the claim that Trump's approval rating for all US citizens is 45%

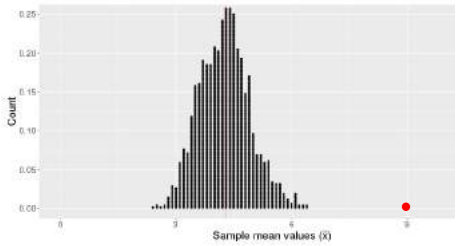
Example 2: we might make the claim that the average height of a baseball player is 72 inches

Basic hypothesis test logic

We start with a claim about a population parameter

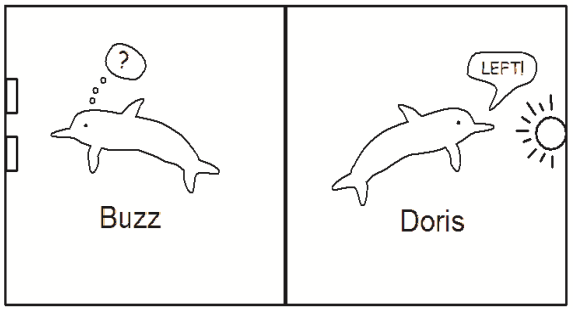
- E.g., $\mu = 4.2$

This claim implies we should get a certain distribution of statistics



If our observed statistic is highly unlikely, we reject the claim

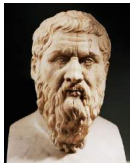
Are dolphins capable of abstract communication?



- Stead light = push button on right
- Flashing = push button on the left

Buzz got 15 out of 16 trials correct

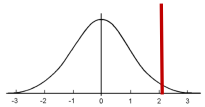
Five steps of hypothesis testing



1. State H_0 and H_A
   - Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)



4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value

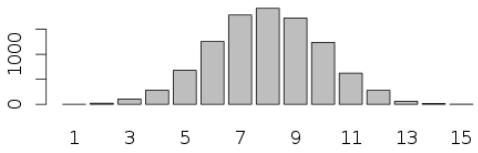
5. Make a judgement
   - Assess whether the results are statistically significant

Steps hypothesis testing – Doris and Buzz

1. State the null hypothesis… and the alternative hypothesis
   - Buzz is just guessing so the results are due to chance: H_0: $\pi = 0.5$
   - Buzz is getting more correct results than expected by chance: H_A: $\pi > 0.5$
   - Rules of the game: $\alpha < 0.05$
2. Calculate the observed statistic
   - Buzz got 15 out of 16 guesses correct, or $\hat{p} = .973$
3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Buzz was just guessing

Simulating Flipping 16 coins 10,000

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0



A **null distribution** is the distribution of statistics one would expect if the null hypothesis (H_0) was true

flip_simulations <- rbinom(num_sims, size, prob)

Steps hypothesis testing – Doris and Buzz

1. State the null hypothesis… and the alternative hypothesis
   - Buzz is just guessing so the results are due to chance: H_0: $\pi = 0.5$
   - Buzz is getting more correct results than expected by chance: H_A: $\pi > 0.5$
   - Rules of the game: $\alpha < 0.05$
2. Calculate the observed statistic
   - Buzz got 15 out of 16 guesses correct, or $\hat{p} = .973$
3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Buzz was just guessing
4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that the dolphins would guess 15 or more correct?
   - i.e., what is the p-value

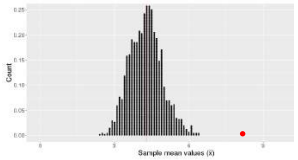
P-values

A **p-value** is the probability, of obtaining a statistic as as (or more) extreme than the observed sample *if the null hypothesis was true*

- i.e., the probability that we would get a statistic as extreme as our observed statistic from the null distribution

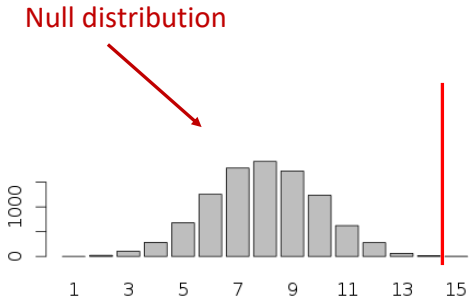
$$\Pr(\text{STAT} \geq \text{observed statistic} \mid H_0 = \text{True})$$

The smaller the p-value, the stronger the statistic evidence is against the null hypothesis



Buzz and Doris example

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0



p-value = 3/10000 = .0003

Steps hypothesis testing – Doris and Buzz

1. State the null hypothesis… and the alternative hypothesis
   - Buzz is just guessing so the results are due to chance: H_0: $\pi = 0.5$
   - Buzz is getting more correct results than expected by chance: H_A: $\pi > 0.5$
   - Rules of the game: $\alpha < 0.05$
2. Calculate the observed statistic
   - Buzz got 15 out of 16 guesses correct, or $\hat{p} = .973$
3. Create a null distribution that is consistent with the null hypothesis
   - i.e., what statistics would we expect if Buzz was just guessing
4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that the dolphins would guess 15 or more correct?
   - i.e., what is the p-value
5. Make a judgement
   - If we have a small p-value, this means that $\pi = .5$ is unlikely and so $\pi > .5$
   - i.e., we say our results are 'statistically significant'

Statistical significance

When our observed sample statistic is unlikely to come from the null distribution, people often say the results are **statistically significant**

- i.e., our p-value is less than α

'Statistically significant' results mean we have strong evidence against H_0 in favor of H_a

- The American Statistical Association rejects the phrase 'statistically significant' significant'

"Matthews ... we're getting another one of those strange 'aw blah es span yol' sounds."

Is it possible to smell whether someone has Parkinson's disease?

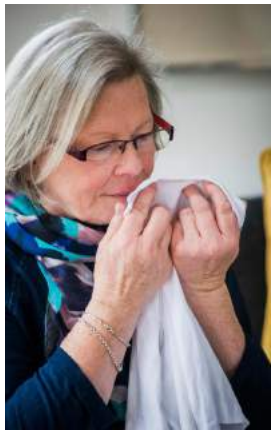
Joy Milne claimed to have the ability to smell whether someone had Parkinson's disease

To test this claim researchers gave Joy 6 shirts that had been worn by people who had Parkinson's disease and 6 people who did not

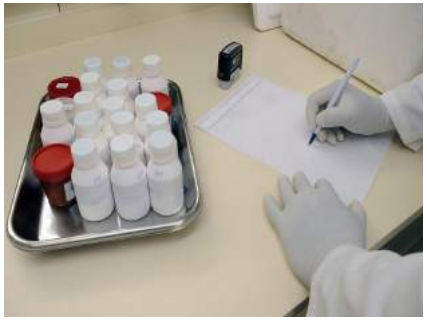
Joy identified 11 out of the 12 shirts correctly

Question: Can Joy really smell whether someone has Parkinson's disease?

Let's examine this in R…



Hypothesis tests for two means



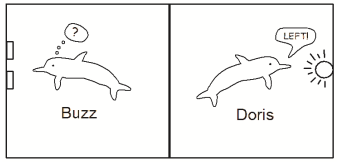
Overview

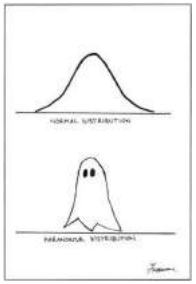
One-tailed and two-tailed tests

Hypothesis tests for two means

Hypothesis tests for more than 2 means

One-tailed vs. two-tailed tests





In the examples we have seen, we were just interested if the parameter was **greater** than an hypothesized parameter

$H_0: \pi = 0.25$ $H_A: \pi > 0.25$

In other cases we might not have a directional alternative hypothesis

Testing whether a coin is biased

Suppose we wanted to test what whether Buzz chose the correct food well more ***or less*** than 50% of the time

- e.g., Buzz might not like the food so was avoiding the well with the food

1. Write down the null and alternative hypotheses

$H_0: \pi = 0.5$

$H_A: \pi \neq 0.5$

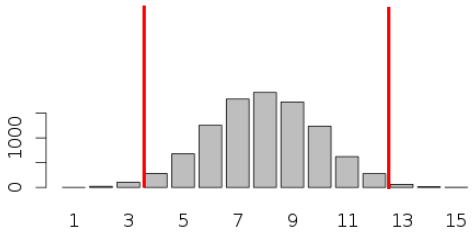
2. Suppose out of the 16 trials, Buzz got the correct 3 times. How would we use a randomized distribution to tell if the coin is biased?

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0

2. Suppose out of the 16 trials, Buzz got the correct 3 times. How would we use our randomized distribution to tell?

3. Based on this table, what is the p-value?

0	0
1	1
2	22
3	105
4	283
5	679
6	1257
7	1786
8	1920
9	1726
10	1238
11	623
12	279
13	63
14	15
15	3
16	0



p-value = 209/10000 = .0209

Compare this p-value to we would have gotten if we **expected** Buzz to avoid the food well?

Statement of alternative hypothesis is important

We need to state what you expect before analyzing the data

Our expectation (hypothesis statement) can change the p-value

Suppose I pulled these 4 cards from a deck

- Probability = 1/52 x 1/51 x 1/50 x 1/49 = 1.5 x 10^{-7}
- Impressed?

What if I told you I was going to pull those exact cards before I did it?

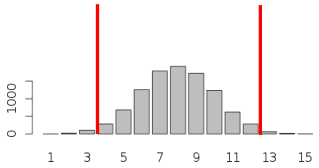
Estimating a p-value from a randomized distribution

For a one tailed alternative: Find the proportion of randomized samples that equal or exceed the original statistic in the direction (tail) indicated by the alternative hypothesis

For a two-tailed alternative: Find the proportion of randomization samples in the tails beyond the observed statistic and 1 - the observed statistic

- Alternatively, find the proportion of randomization samples in the smaller tail at or beyond the original statistic and then double the proportion to account for the other tail

How to estimate two sided p-values in R?




```
null_dist <- rbinom(10000, 16, .5)    # numbers 0 to 16
null_dist <- null_dist/16   # convert to p̂ numbers from 0 to 1

num_right_tail  <-  sum(null_dist >= 13/16)
num_left_tail  <-  sum(null_dist <= 3/16)

p_value <- (num_right_tail + num_left_tail)/10000
```

Inference on a single proportion: Paul the Octopus

In the 2010 World Cup, Paul the Octopus (in a German aquarium) became famous for correctly predicting 11 out of 13 soccer games.

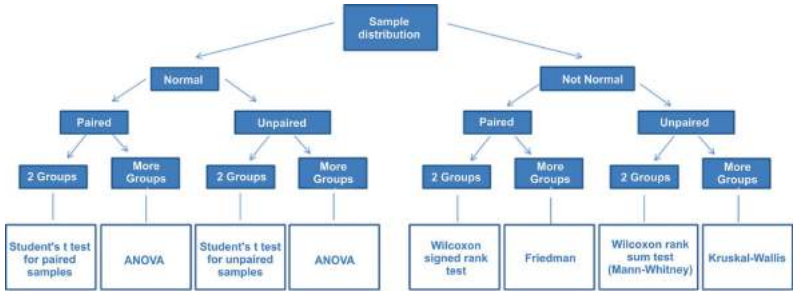
Question: is Paul psychic?

Homework 4!

Before we start: the big picture…

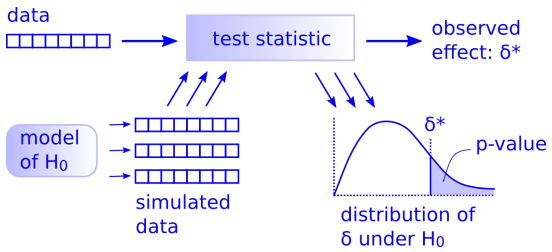
There are many different parameters we might want to test:

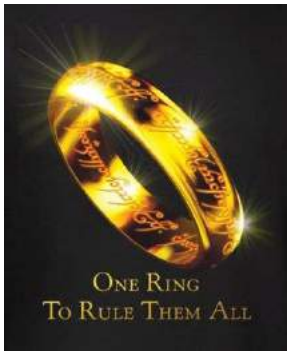
π, μ, β, ρ, $\mu_1 = \mu_2$, etc.



Before we start: the big picture…

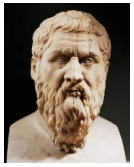
There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

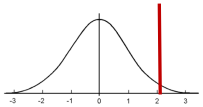
Five steps of hypothesis testing



1. State H_0 and H_A
   - Assume Gorgias (H_0) was right

2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)



4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value

5. Make a judgement
   - Assess whether the results are statistically significant

Hypothesis tests for comparing two means

Question: Is this pill effective?

Testing whether a pill is effective

How would we design a study?

What would the cases and variables be?

What would the parameter and statistic of interest be?

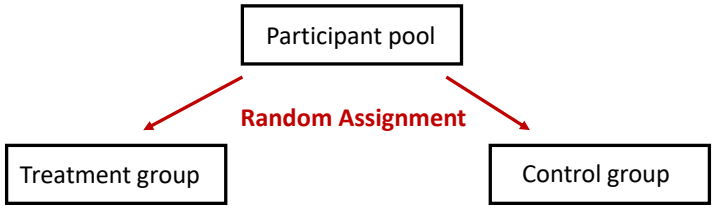
What are the null and alternative hypotheses?

- Assume we are looking for differences in means between the groups

Experimental design

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get the pill
- Half in a *control group* where they get a fake pill (placebo)
- See if there is more improvement in the treatment group compared to the control group



Hypothesis tests for differences in two group means

1) State the null and alternative hypothesis
   - H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
   - H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} > 0$

2) Calculate statistic of interest
   - $\bar{x}_{Effect} = \bar{x}_{Treatment} - \bar{x}_{Control}$

Example: Does calcium reduce blood pressure?

A randomized by Lyle et al (1987) comparative experiment investigated whether calcium lowered blood pressure in African-American men

- A treatment group of 10 men received a calcium supplement for 12 weeks
- A control group of 11 men received a placebo during the same period

The blood pressure of these men was taken before and after the 12 weeks of the study

1) What are the null and alternative hypotheses?
   - H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
   - H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} > 0$
     - i.e., a greater decrease in blood pressure after taking calcium

Does calcium reduce blood pressure?

Treatment data (n = 10):

Begin	107	110	123	129	112	111	107	112	136	102
End	100	114	105	112	115	116	106	102	125	104
Decrease	**7**	**-4**	**18**	**17**	**-3**	**-5**	**1**	**10**	**11**	**-2**

Control data (n = 11):

Begin	123	109	112	102	98	114	119	112	110	117	130
End	124	97	113	105	95	119	114	114	121	118	133
Decrease	**-1**	**12**	**-1**	**-3**	**3**	**-5**	**5**	**2**	**-11**	**-1**	**-3**

2) What is the observed statistic of interest?
- $\bar{x}_{Effect}$ = 5 - -.2727 = 5.273

3) What is step 3?

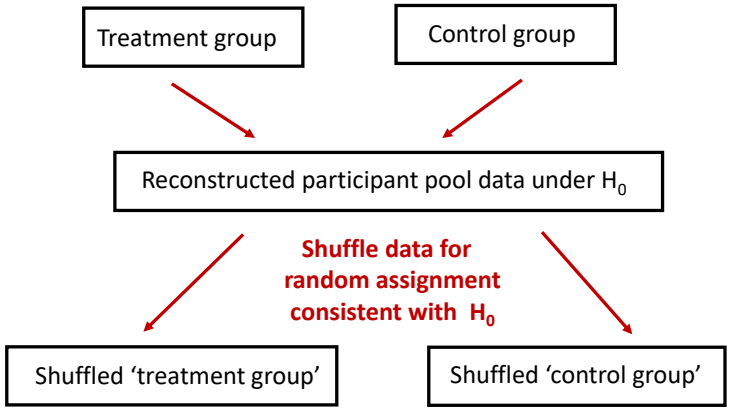
3. Create the null distribution!

How could we create the null distribution?

Need to generate data consistent with H_0: $\mu_{Treatment} - \mu_{Control} = 0$
- i.e., we need fake $\bar{x}_{Effect}$ that are consistent with H_0

Any ideas how we could do this?

3. Create the null distribution!



One null distribution statistic: $\bar{x}_{Shuff\_Treatment} - \bar{x}_{Shuff\_control}$

3. Create a null distribution

1) Combine data from both groups

2) Shuffle data

3) Randomly select 10 points to be the 'null' treatment group

4) Take the remaining points to the 'null' control group.

5) Compute the statistic of interest on these 'null' groups

6) Repeat 10,000 times to get a null distribution

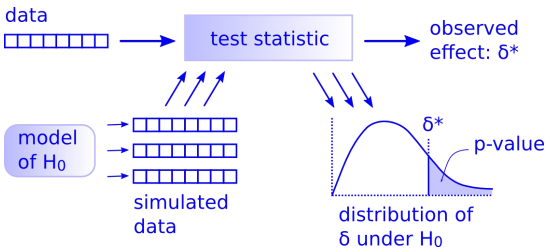
Let's try the rest of the hypothesis test in R…

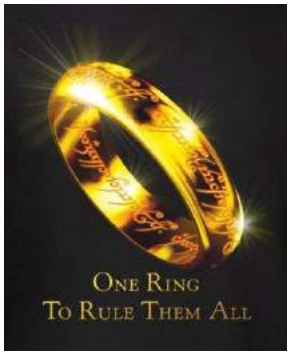
Hypothesis tests for more than two means

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Before we start: the big picture…

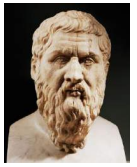
There is only one [hypothesis test]():





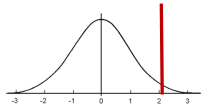
Just follow the 5 hypothesis tests steps!

Five steps of hypothesis testing



1. State H_0 and H_A
   - Assume Gorgias (H_0) was right
   - α = .05 of the time he will be right, but we will say he is wrong
2. Calculate the actual observed statistic

3. Create a distribution of what statistics would look like if Gorgias is right
   - Create the **null distribution** (that is consistent with H_0)
4. Get the probability we would get a statistic more than the observed statistic from the null distribution
   - p-value



5. Make a judgement
   - Assess whether the results are statistically significant

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	

Comparing more than two means

A group of Hope College students wanted to see if there was an association between a student's major and the time it takes to complete a small Sudoku-like puzzle

They grouped majors into four categories

- Applied science (as)
- Natural science (ns)
- Social science (ss)
- Arts/humanities (ah)

What is the first thing to do to analyze the data?

Sudoku by field

1. State the null and alternative hypotheses!

What should we do next?

Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:
   $\max \bar{x} - \min \bar{x}$
2. Mean absolute difference (MAD):
   $(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$
3. F statistic:

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Using the MAD statistic

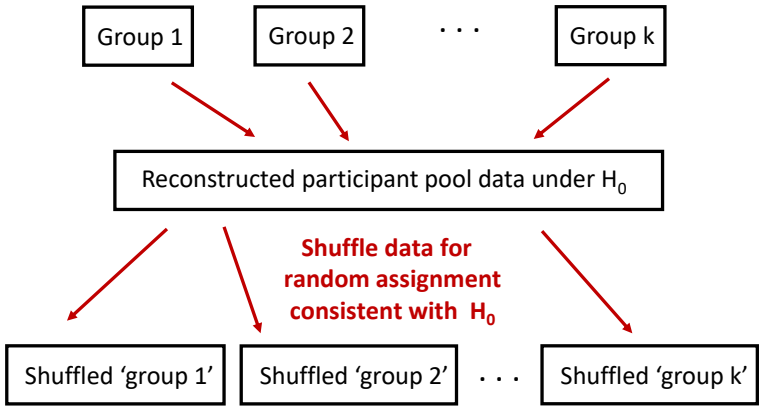
Mean absolute difference (MAD):

$$(|\bar{x}_{as} - \bar{x}_{ns}| + |\bar{x}_{as} - \bar{x}_{ss}| + |\bar{x}_{as} - \bar{x}_{ah}| + |\bar{x}_{ns} - \bar{x}_{ss}| + |\bar{x}_{ns} - \bar{x}_{ah}| + |\bar{x}_{ss} - \bar{x}_{ah}|)/6$$

Observed statistic value = 13.92

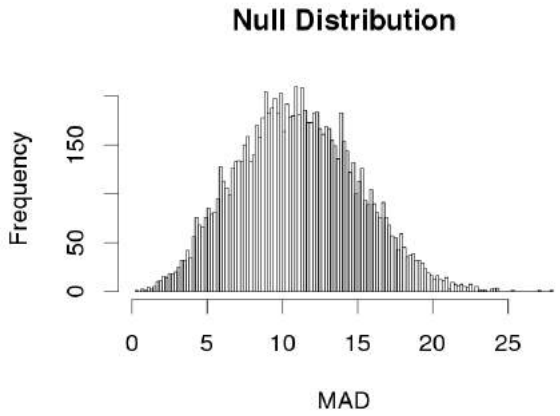
How can we create the null distribution?

3. Create the null distribution!

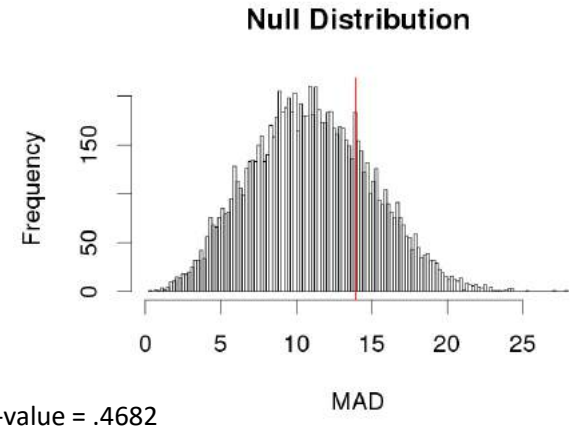


Compute statistics from shuffled groups

Null distribution



P-value



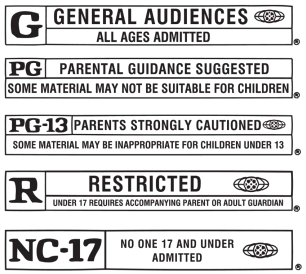
p-value = .4682

Conclusions?

Let's start on trying this analysis in R…

Homework 5

Rotten Tomatoes is a website that provides movie ratings and reviews



Question: Do critics' rate movies the same on average regardless of their MPAA ratings?

z-scores and correlation

z-scores

The z-scores tells how many standard deviations a value is from the mean mean

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

Which statistic is most impressive?

League statistics:

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

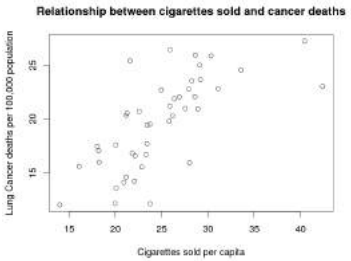
League statistics:

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

The summary statistics of the NBA in 2011 are given below

$$z = (x - \bar{x}) / s$$

Correlation

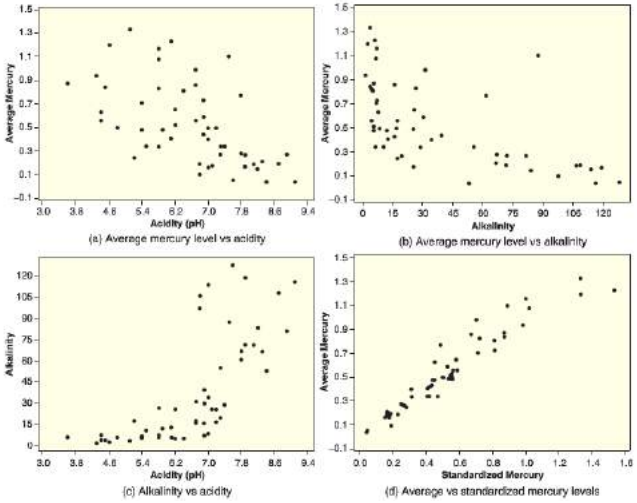


The **correlation coefficient** is a statistic measure of the strength and direction of a linear association between two variables

- Correlation as always between -1 and 1: $-1 \le r \le 1$
- The sign of r indicates the direction of the association
- Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Florida lakes

Correlation game



The correlation coefficient

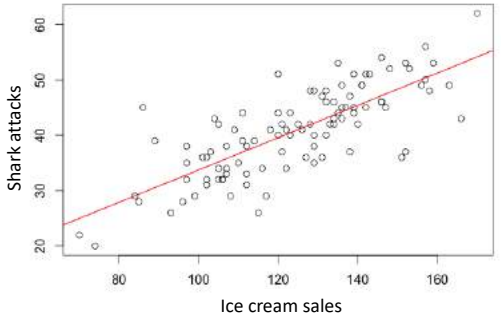
The correlation coefficient statistic (r) can be calculated as:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

R: `cor(x, y)`

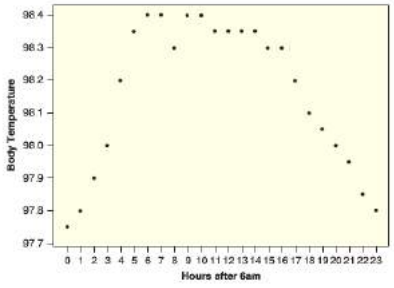
Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply cause and effect relationship between two variables



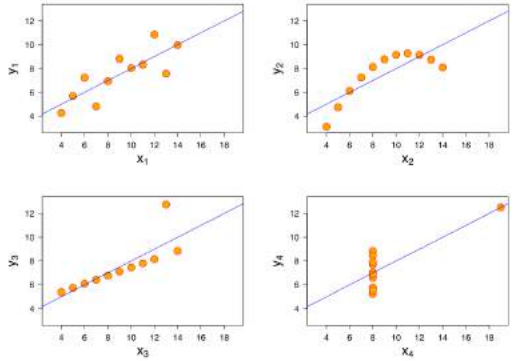
Correlation caution #2

A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

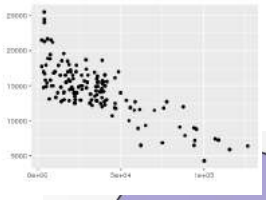


Correlation caution #3

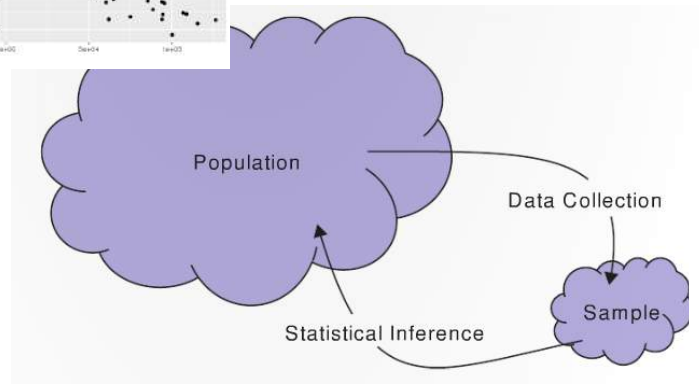
Correlation can be heavily influences by outliers. Always plot your data!



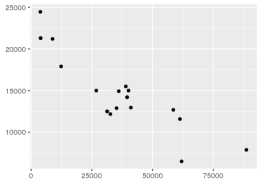
Anscombe's quartet
(r = 0.81)



ρ parameter



r statistic



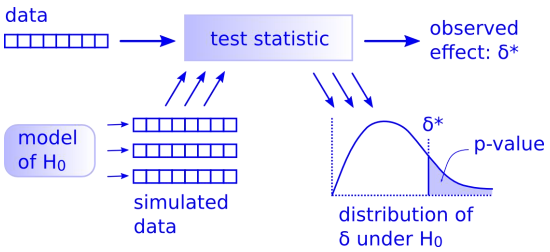
Hypothesis tests for correlation

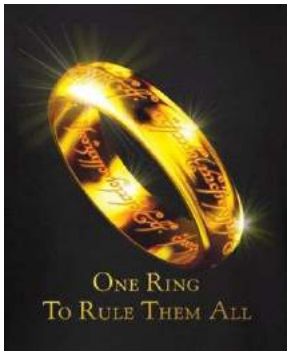
Suppose we wanted to test whether there was a positive correlation between two variables

How could we write this using symbols?

One test to rule them all

There is only one hypothesis test!





Just follow the 5 hypothesis tests steps!

Hypothesis tests for correlation

Question: Is there a positive correlation between the number of carbohydrates in a cereal and the number calories?

How could we go about answering this question?

Significance tests for correlation

Suppose we had some data from 30 randomly selected cereals

	Calories	Carbohydrates
AppleJacks	117	27
Boo Berry	118	27
Cap'n Crunch	144	31
Cinnamon Toast Crunch	169	32

What would be a good first step in analyzing the data?

Let's try it in R…

1969 Vietnam Draft

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The first date picked was Sept 14 (sequential number 258)

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The second date picked was April 24th (sequential number 115)

What is your Draft number?

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

1969 Vietnam Draft sorted by sequential date

Date	Sequential date	Draft number
Jan 1	1	305
Jan 2	2	159
Jan 3	3	251
Jan 4	4	215
Jan 5	5	101
Jan 6	6	224
Jan 7	7	306
Jan 8	8	199
Jan 9	9	194

1969 Vietnam Draft

In a perfectly fair, random lottery, what should be the value of the correlation coefficient between **draft number** and **sequential date** of birthday?

Homework 5

Test if the 1969 Vietnam draft lottery was completely random

- i.e., no correlation between sequential data and draft number

Hypothesis tests for correlation, theories of hypothesis tests, and parametric tests

Overview

Correlation and z-scores

Permutation tests for correlation

Theories of hypothesis testing

Parametric hypothesis tests

z-scores and correlation

z-scores

The z-scores tells how many standard deviations a value is from the mean

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

Which statistic is most impressive?

League statistics:

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

Which Accomplishment is most impressive?

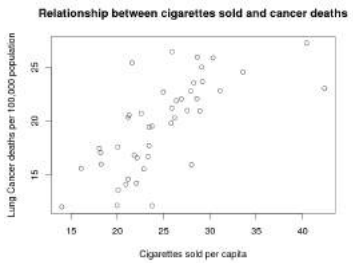
LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

The summary statistics of the NBA in 2011 are given below

z	=	$(x - \bar{x}) / s$		
Z-score FGPct	=	(0.510 - 0.464)/0.053	=	0.868
Z- score Points	=	(2111 - 994)/414	=	2.698
Z-score Assists	=	(554 - 220)/170	=	1.965
Z-score Steals	=	(124 - 68.2)/31.5	=	1.771

Correlation

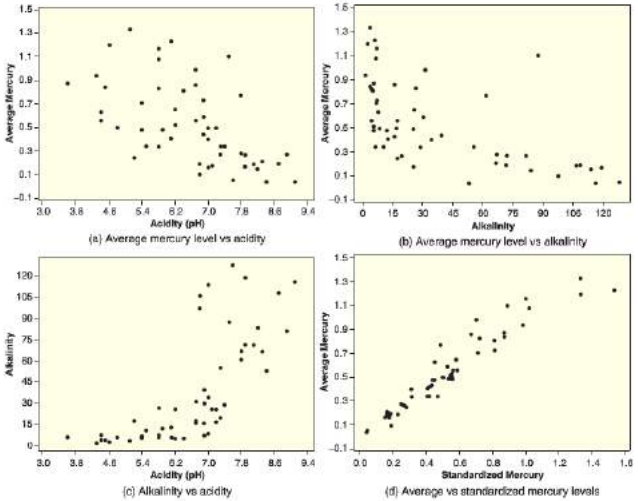


The **correlation coefficient** is a statistic measure of the strength and direction of a <u>linear association</u> between two variables

- Correlation as always between -1 and 1: $-1 \le r \le 1$
- The sign of r indicates the direction of the association
- Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Florida lakes

[Correlation game]



The correlation coefficient

The correlation coefficient statistic (r) can be calculated as:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

R: `cor(x, y)`

Correlation caution #1

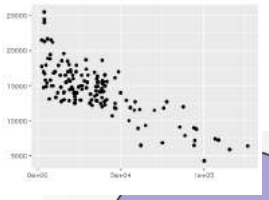
A strong positive or negative correlation does not (necessarily) imply cause and effect relationship between two variables

Correlation caution #2

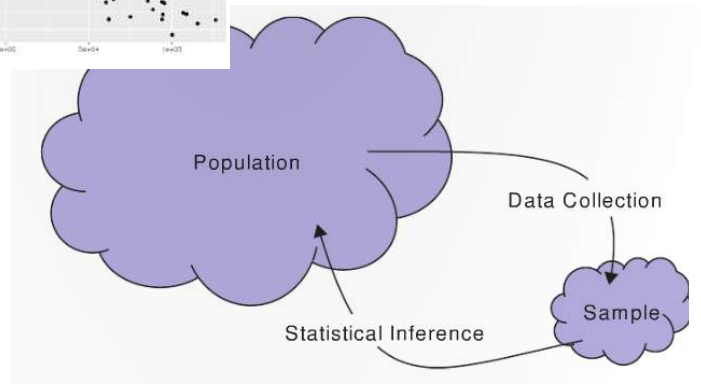
A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Correlation caution #3

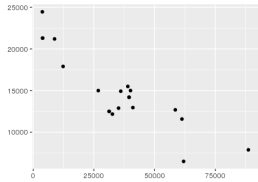
Correlation can be heavily influences by outliers. Always plot your data!



ρ parameter



r statistic



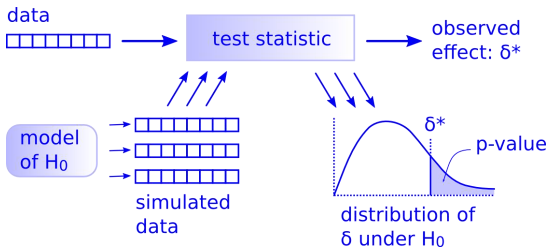
Hypothesis tests for correlation

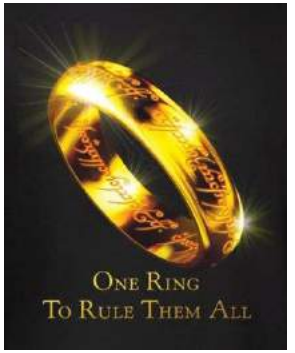
Suppose we wanted to test whether there was a positive correlation between two variables

How could we write this using symbols?

One test to rule them all

There is only one [hypothesis test](#)!





Just follow the 5 hypothesis tests steps!

Hypothesis tests for correlation

Question: Is there a positive correlation between the number of carbohydrates in a cereal and the number calories?

How could we go about answering this question?

Significance tests for correlation

Suppose we had some data from 30 randomly selected cereals

	Calories	Carbohydrates
AppleJacks	117	27
Boo Berry	118	27
Cap'n Crunch	144	31
Cinnamon Toast Crunch	169	32

What would be a good first step in analyzing the data?

Let's try it in R...

1969 Vietnam Draft

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The first date picked was Sept 14 (sequential number 258)

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The second date picked was April 24th (sequential number 115)

What is your Draft number?

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

1969 Vietnam Draft sorted by sequential date

Date	Sequential date	Draft number
Jan 1	1	305
Jan 2	2	159
Jan 3	3	251
Jan 4	4	215
Jan 5	5	101
Jan 6	6	224
Jan 7	7	306
Jan 8	8	199
Jan 9	9	194

1969 Vietnam Draft

In a perfectly fair, random lottery, what should be the value of the correlation coefficient between **draft number** and **sequential date** of birthday?

Homework 5

Test if the 1969 Vietnam draft lottery was completely random

- i.e., no correlation between sequential data and draft number

Two theories of hypothesis testing

Null-hypothesis significance testing (NHST) is a hybrid of two theories:

1. Significance testing of Ronald Fisher
2. Hypothesis testing of Jezy Neyman and Egon Pearson

Fisher (1890-1962)

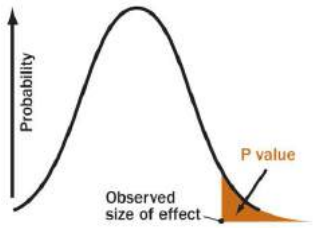
Neyman (1894-1981)

Pearson (1895-1980)

Ronald Fisher's significance testing

Views the p-value as strength of evidence against the null hypothesis

- P-values part of an on-going scientific process: tells the experimenter "what results to ignore"



Neyman-Pearson null hypothesis testing

Makes ***a formal decision*** in statistical tests

Reject H_0: if the observed sample statistic is so extreme is unlikely when H_0 is true

- i.e., reject H0 if the p-value is less than some predetermined **significance level** α

Do not reject H_0: if the statistic is not too extreme when H_0 is true. This means the test is inconclusive.

Frequentist logic

Type I error: incorrectly rejecting the null hypothesis when it is true

If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, then only ~5% of all published research findings should be wrong (for $\alpha = 0.05$)

- i.e., we would only make type I errors 5% of the time

Frequentist logic

Type 2 error: incorrectly rejecting failing to reject H_0 when it is false

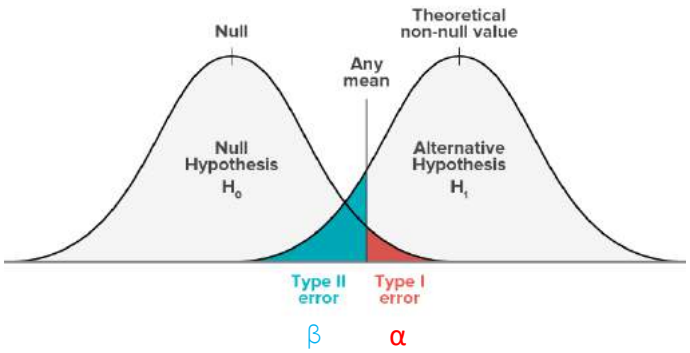
- The rate at which we make type 2 errors is often denoted with the symbol β

The **power** of a test is the probability we reject the H_0 when it is false

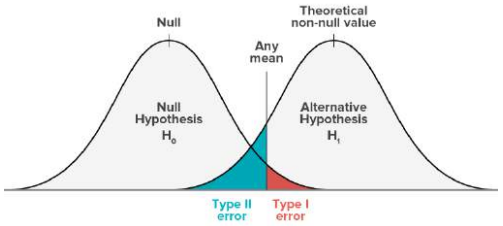
- $1 - \beta$

For a fixed α level, it would be best to use the most powerful test

Type I and Type II Errors



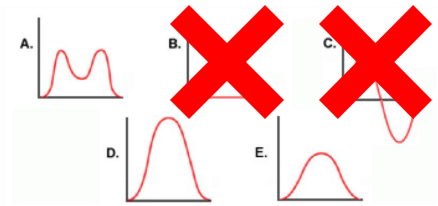
Type I and Type II Errors

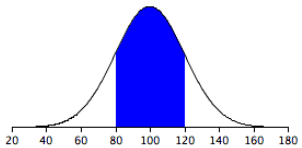


	Reject H_0	**Do not reject H_0**
H_0 is true	Type I error (α) (false positive)	No error
H_0 is false	No error	Type II error (β) (false negative)

Parametric hypothesis tests

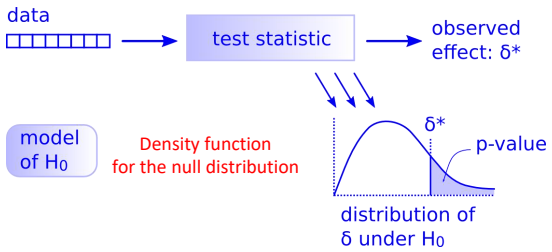
Parametric hypothesis tests are hypothesis test that use density functions for their null distribution

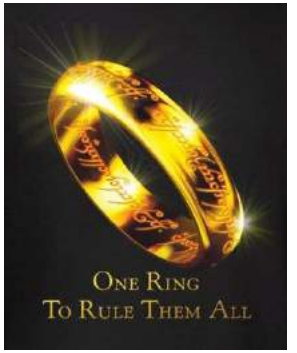




One test to rule them all

There is only one [hypothesis test](#)!





Just follow the 5 hypothesis tests steps!

Parametric hypothesis tests

In order to have a null distribution that is a density function we have to make some **assumptions**

For example, we might have to assume that our statistic comes from a *normal distribution*

Suppose we are looking at the statistic $\bar{x}$

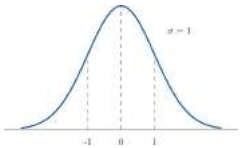
- i.e., doing a test that involves μ

Would assuming out statistic comes from a normal distribution be reasonable?

z statistics

We also often convert these statistics to the standard normal distribution N(0, 1) using a z-transformation:

$$z = \frac{Sample\ \ Statistic\ -\ Null\ \ Parameter}{SE}$$



In this example, the 'Null Parameter' is a hypothesized value

- For example: $\mu = 72$ or $\mu_1 - \mu_2 = 0$

The SE is given by a formula, such as: $SE = \frac{\sigma}{\sqrt{n}}$

Is it possible to use this formula with real data?

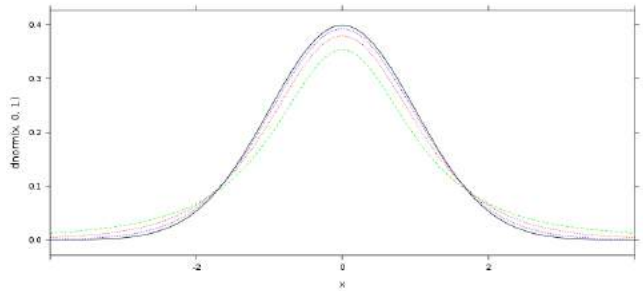
z statistics

$$z = \frac{Sample\ \ Statistic\ -\ Null\ \ Parameter}{SE} \qquad SE = \frac{\sigma}{\sqrt{n}}$$

We can replace σ with s to get: $\hat{SE} = \frac{s}{\sqrt{n}}$

Is it possible to use this formula with real data?

t-distributions



N(0, 1), df = 2, df = 5, df = 15

This can be a useful model for our null distribution!

t-test for single mean

To test:

H_0: $\mu = \mu_0$ vs.

H_A: $\mu \neq \mu_0$ (or a one-tailed alternative)

We use the t-statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

A p-value can be computed using a t-distribution with n-1 degrees of freedom

- Provided that the population is reasonable normal (or the sample size is large)

The Chips Ahoy! Challenge

In the mid-1990s a Nabisco marking campaign claimed that there were at least 1000 chips in every bag of Chips Ahoy! cookies.

A group of Air Force cadets tested this claim by dissolving the cookies from 42 bags in water and counting the number of chips.

They found the average number of chips per bag was 1261.6, with a standard deviation of 117.6 chips

Test whether the average number of chips per bag is greater than 1000. Do the results confirm Nabisco's claim?

$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ pt(t, df = deg_of_free)

The Chips Ahoy! Challenge

Let's try the analysis in R!

Parametric tests

Overview

Information about midterm and final exam/project

Quick review: theories of hypothesis testing

Parametric hypothesis tests
- For one mean
- For two means

Simple linear regression (if there is time)

Midterm exam

In class on October 24th – 15% of your grade
- TAs will have review session during class on October 22nd

Format
- Multiple choice questions, short answers, review of code
- Write a short amounts of code

Topics covered
- Everything we discussed in class
  - Confidence intervals, hypothesis tests, theories of hypothesis tests, etc.

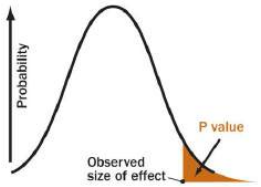
Final exam

In class on December 14 at 7pm

Was planning on splitting in into a final project and final exam
- Final exam 15% of your grade, similar to the midterm
- Final project 10% of your grade, ~5-8 page report

Review: two theories of hypothesis testing

Significance testing
(Fisher)



p-value is strength of evidence against H_0

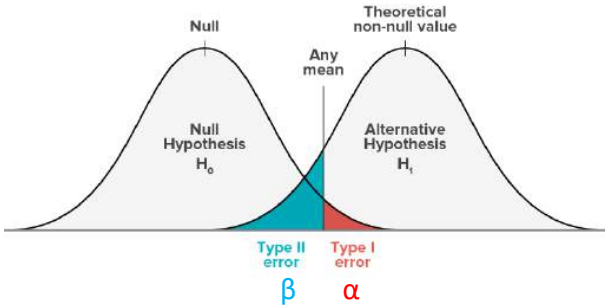
"inductive inference"

Hypothesis testing
(Neyman and Pearson)

Proportion of times incorrectly rejecting H_0 is controlled for

"inductive behavior"

Neyman and Pearson paradigm: two types of errors



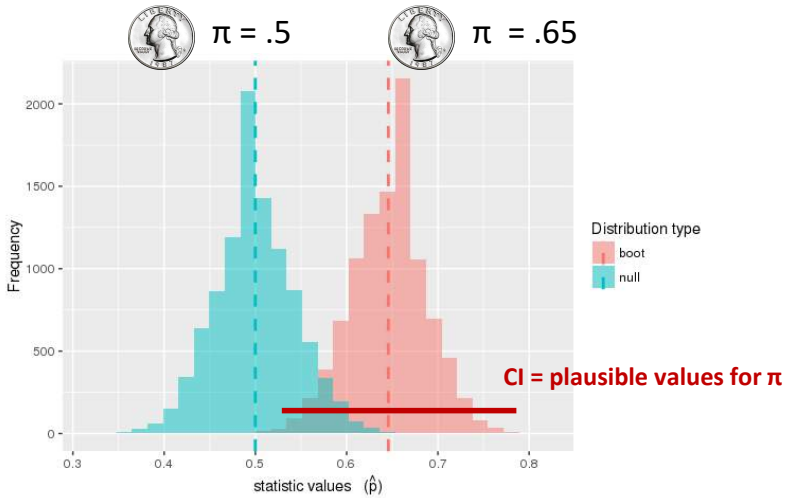
If we set α = 0.05 and the H_0 is true

Then we should falsely reject H_0 ~5% of the time

Bootstrap vs. null distribution

Suppose our CI does not contain the parameter specified by H_0

Then we can reject H_0



Problems with the NP hypothesis tests

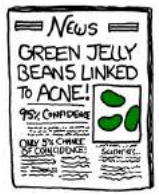
Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., So what if 95% of literature is true if we are wrong

Problem 2: Arbitrary thresholds for alpha levels

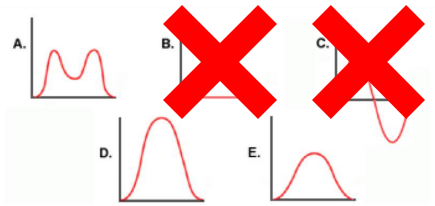
- P-value = 0.051, we don't reject H_0?

Problem 3: running many tests can give rise to a high number of type 1 errors

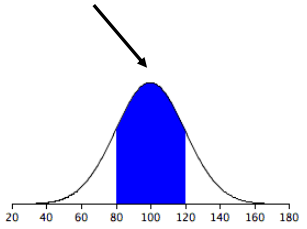


Parametric hypothesis tests

Parametric hypothesis tests are hypothesis test that use **density functions** for the **null distribution**

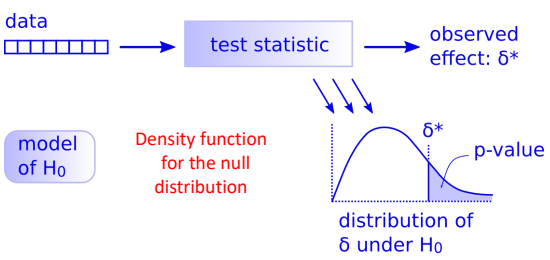


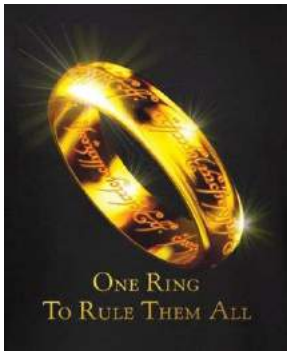
Parametric tests:
When H_0 is true, this is the null distribution



One test to rule them all

There is only one [hypothesis test](#)!



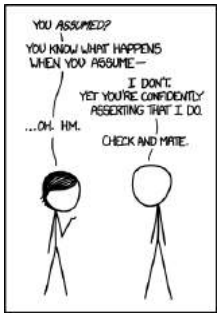


Just follow the 5 hypothesis tests steps!

Parametric hypothesis tests

In order to have a null distribution that is a density function we have to make some (additional) **assumptions**

- These tests are based on deductive logic proofs where we start with set of axioms and derive conclusions
- If the axioms are valid for our real data, these tests give valid results
  - If the axioms/assumptions are not true, they could give incorrect results
- It is important to check to see the assumptions appear valid for our data by:
  - Looking at visual plots of our data
  - Running additional tests to see if the assumptions are met
    - (although often these tests are weak so visual inspection can be better)



Parametric hypothesis tests

An example of an assumption: we often assume that our statistic (and/or data) come from a *normal distribution*

Suppose we are looking at the statistic $\bar{x}$

- i.e., doing a test that involves μ

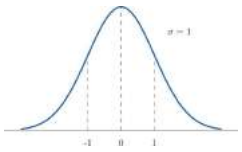
Would assuming out statistic comes from a normal distribution be reasonable?

- What if the sample size was n = 1,000?
  - Based on the Central Limit Theorem, very likely

z statistics

For tests where we assume that data comes from a normal distribution, we often convert our initial statistic into the standard normal distribution N(0, 1) using a z-transformation:

$$z = \frac{Sample\ \ Statistic\ -\ Null\ \ Parameter}{SE}$$



The 'Null Parameter' is the hypothesized parameter value

- For example: $\pi = .5$ or $\mu_1 - \mu_2 = 0$

The SE is given by a formula, such as: $SE = \frac{\sigma}{\sqrt{n}}$

Is it possible to use this formula with real data?

z statistics

$$z = \frac{Sample\ \ Statistic\ \ -\ \ Null\ \ Parameter}{SE} \qquad SE = \frac{\sigma}{\sqrt{n}}$$

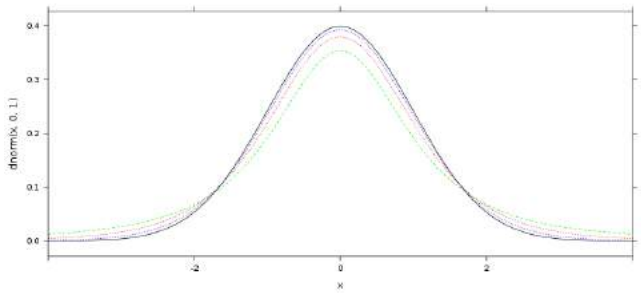
We can replace σ with s to get: $\hat{SE} = \frac{s}{\sqrt{n}}$

Is it possible to use this formula with real data?

Yes! But our statistic now comes from a **t-distribution** instead of a normal distribution

- In particular, a t-distribution with n - 1 degrees of freedom

t-distributions



N(0, 1) df = 2 df = 5 df = 15

Density: dt(x_vals, df)

$P(T > t_{stat})$: pt(t_stat, df, lower.tail = FALSE)

The distribution of sample means using the sample standard deviation

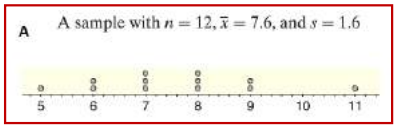
$$\frac{\bar{x}\ -\ \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

A t-distribution with
n - 1 degrees of freedom

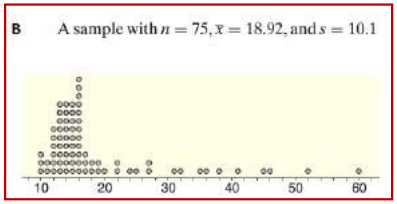
This works if:

- The underlying population has a distribution that is approximately normal
- OR our sample size is greater than 30; i.e., n > 30

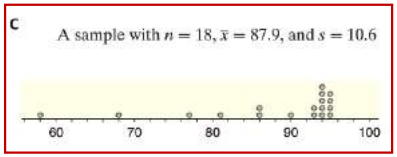
Is the t-distribution appropriate?



Distribution seems normal so OK to use t-distribution



Sample size is larger than n = 30 so OK to use the t-distribution



Population distribution does not look normal and n < 30 so NOT ok to use the t-distribution

t-test for single mean

Suppose we wanted to test:

H_0: $\mu = \mu_0$ vs.

H_A: $\mu \neq \mu_0$ (or a one-tailed alternative)

We can use the t-statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

A p-value can be computed using a t-distribution with n-1 degrees of freedom

- i.e., we can use the pt(t, df) function to get the p-value
- The inference is valid if the population is reasonable normal or the sample size is large

The Chips Ahoy! Challenge

In the mid-1990s a Nabisco marking campaign claimed that there were at least 1000 chips in every bag of Chips Ahoy! cookies.

A group of Air Force cadets tested this claim by dissolving the cookies from 42 bags in water and counting the number of chips.

They found the average number of chips per bag was 1261.6, with a standard deviation of 117.6 chips

Test whether the average number of chips per bag is greater than 1000. Do the results confirm Nabisco's claim?

$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$ pt(t, df = deg_of_free)

Let's try it in R...

John Tukey quote

"Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question"

- *The future of data analysis*. Annals of Mathematical Statistics 33 (1), (1962), page 13.

Distribution of differences in means

We can also use t-distributions for comparing 2 means

- i.e., we could test: H_0: $\mu_1 - \mu_2 = 0$, vs. H_A: $\mu_1 - \mu_2 \neq 0$
- Suppose we have two samples of size n_1 and n_2

$$t = \frac{stat - null\ param}{\hat{SE}}$$

The formula for our standard error is: $\hat{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

We can then calculate the t-statistic using: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

We can use the t-distribution if the sample is large (>30) or if the population is reasonably normal

We can use the df as the smaller of $n_1 - 1$ or $n_1 - 1$

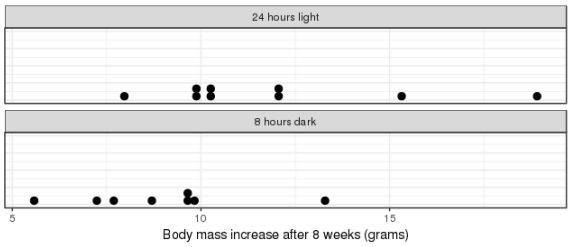
- or technology to get a better approximation of the df

Re-examine: do mice who eat late at night get fat?

A study by Fonken et al, 2010, wanted to examine whether more weight was gained by mice who could eat late at night

Mice were randomly divided into 2 groups:

- Dark condition: 8 mice were given 8 hours of darkness at night (when they couldn't eat)
- Light condition: 9 were constantly exposed to light for 24 hours (so they could always eat)



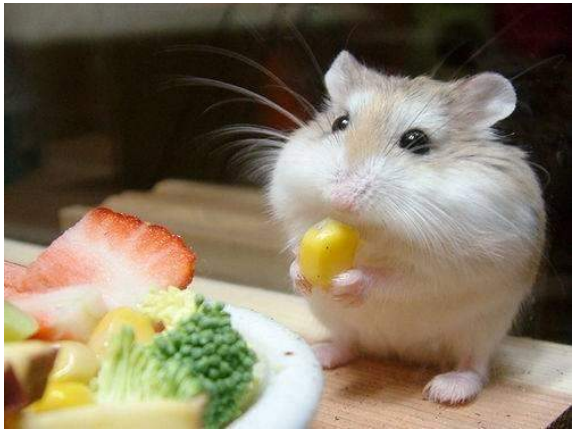
Hypothesis tests for differences in two group means

1. State the null and alternative hypothesis
   - H_0: $\mu_{Light} = \mu_{Dark}$ or $\mu_{Light} - \mu_{Dark} = 0$
   - H_A: $\mu_{Light} > \mu_{Dark}$ or $\mu_{Light} - \mu_{Dark} > 0$

2. Calculate statistic of interest: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

3-5. Use the pt() function to find the p-value and make a decision

Let's try it in R



This mouse eats healthy!

Parametric tests

Overview

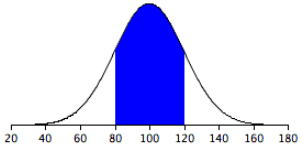
Parametric hypothesis tests for more than 2 means (ANOVA)

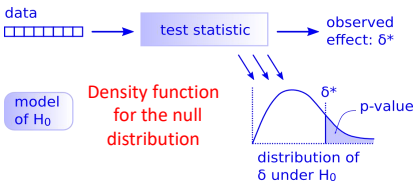
Simple linear regression

Parametric hypothesis tests

Parametric hypothesis tests are hypothesis test that use **density functions** for the **null distribution**

- i.e., by making some additional assumptions, we can know the mathematical form of the null distribution





We need to make sure to check our assumptions to make sure these parametric tests are valid in practice

The t-distribution is a common parametric null distribution

Similar to a Z statistic from N(0, 1) but we are using s instead of σ

$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

A t-distribution with n - 1 degrees of freedom

This works if:

- The underlying population has a distribution that is approximately normal
- OR our sample size is greater than 30; i.e., n > 30

t-test for single mean

Suppose we wanted to test:

H_0: $\mu = \mu_0$ vs.

H_A: $\mu \neq \mu_0$ (or a one-tailed alternative)

We can use the t-statistic: $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$ $\quad t = \frac{stat - null\ param}{\hat{SE}}$

A p-value can be computed using a t-distribution with n-1 degrees of freedom

- i.e., we can use the pt(t, df) function to get the p-value
- The inference is valid if the population is reasonable normal or the sample size is large

t-test for comparing two means

Suppose we wanted to test:

H_0: $\mu_1 - \mu_2 = 0$

H_A: $\mu_1 - \mu_2 \neq 0$ (or a one-tailed alternative)

We can use the t-statistic: $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $\quad t = \frac{stat - null\ param}{\hat{SE}}$

We can use the t-distribution if the sample is large (>30) or if the population is reasonably normal

- We can use the df as the smaller of $n_1 - 1$ or $n_1 - 1$

Parametric test for comparing more than one mean: One-way ANOVA

An Analysis of Variance (ANOVA) is a test that can be used to examine a set of means are all the same

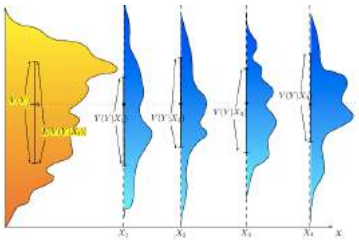
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- $\mu_i \neq \mu_j$ for some i, j

The statistic we use for a one-way ANOVA is the F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

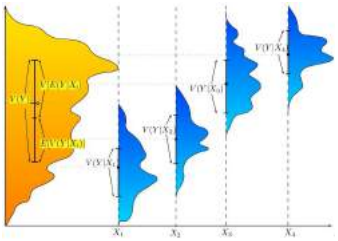
The F-Statistic

If data from all groups came from **the same distribution**



- Similar means
- Similar spread

If data from all groups came from **different distributions**

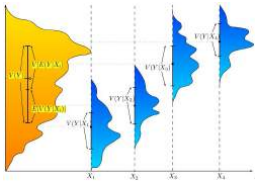


- Different means
- Smaller spreads

The F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

K: the number of groups
n_i: the number of points in group i
x_{ij} :the j^{th} data point from group i
$\bar{x}_i$: the mean of group i
$\bar{x}_{tot}$: the mean across all the data



When the null hypothesis is true, F has a value around 1

- The numerator and denominator are both estimate of σ^2

Parametric test for comparing more than one mean: One-way ANOVA

The F-statistic comes from a F-distribution

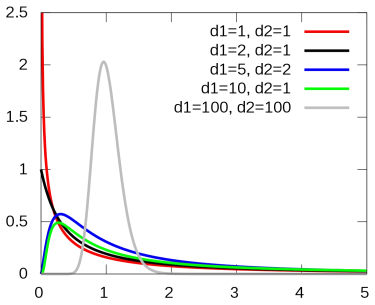
- df1 = K - 1
- df2 = N - K

Assumptions underlying a one-way ANOVA

- Data in each group come from normal distributions
- Each group has equal variance (homoskedasticity)

Can check these assumptions by:

- Visualizing data in each group
- Seeing if the ratio of $s_i/s_j > 2$



One-way ANOVA table

Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{i=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_t)^2$		

k: number of groups
n: number of data points

$$\frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

One-way ANOVA in R

Let's briefly look at how to use R's built in functions to do a one-way ANOVA…

Also, check out this interactive tutorial on applying ANOVAs to neural data created by Brooke Fitzgerald

https://neuraldata.net/

Regression

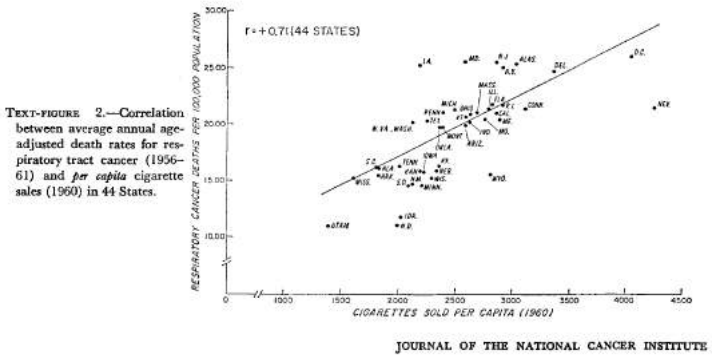
Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

- In simple linear regression, we use a single variable x, to predict y

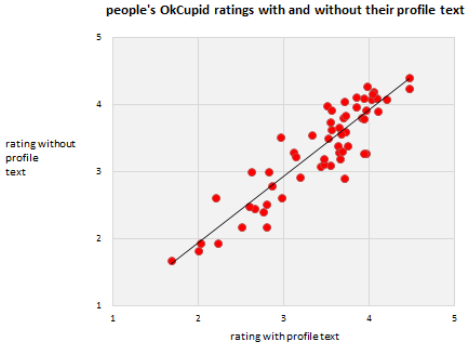
Cigarettes cancer regression line



TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956–61) and *per capita* cigarette sales (1960) in 44 States.

JOURNAL OF THE NATIONAL CANCER INSTITUTE

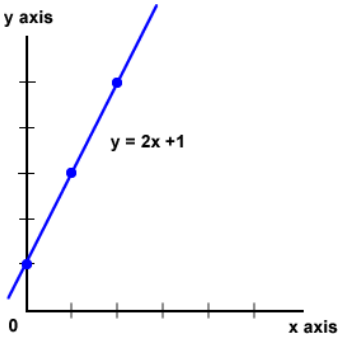
OkCupid text and images



Equation for a line

What is the equation for a line?

$$\hat{y} = b_0 + b_1 \cdot x$$



Regression lines

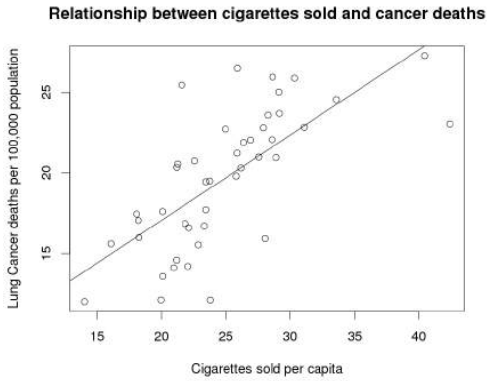
$$\hat{y} = b_0 + b_1 \cdot x$$

$$Response = b_0 + b_1 \cdot Explanatory$$

The slope b_1 represents the predicted change in the response variable y given a one unit change in the explanatory variable x

The intercept (b_0) a represented the predicted value of the response variable y if the explanatory variable x were 0

Cancer smoking regression line



$\hat{y} = b_0 + b_1 \cdot x$

$b_0 = 6.47$

$b_1 = 0.53$

R: `lm(y ~ x)`

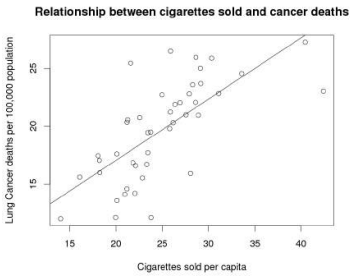
Using the regression line to make predictions

If a state sold 25 cigarettes per person

How many cancer deaths (per 100,000 people) would you expect?

$b_0 = 6.47, \quad b_1 = .53$

$\hat{y} = 6.47 + .53 \cdot x$

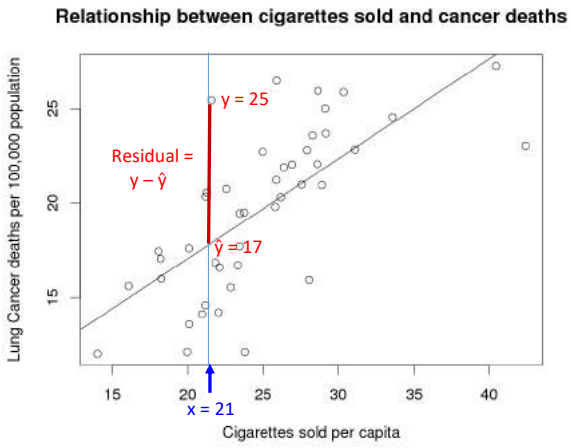


Residuals

The **residual** at a data value is the difference between the observed (y) and predicted value of the response variable

Residual = Observed – Predicted = $y - \hat{y}$

Cancer smoking residuals

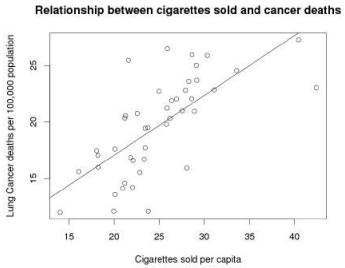


Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)
17.05	16.10	0.95
19.80	20.13	-0.33
15.98	16.12	-0.14
22.07	21.60	0.47
22.83	22.93	-0.10
24.55	24.25	0.30
27.27	27.88	-0.61
23.57	21.24	2.14

Line of 'best fit'

The **least squares line**, also called '**the line of best fit'**, is the line which minimizes the sum of squared residuals



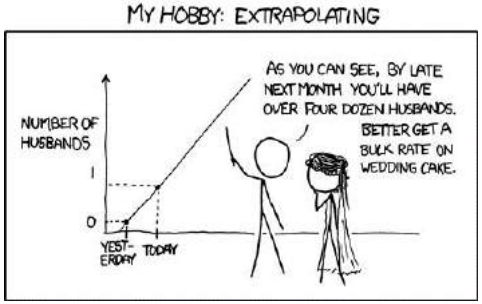
Try to find the line of best fit

Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)	Residuals² (y - ŷ)²
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

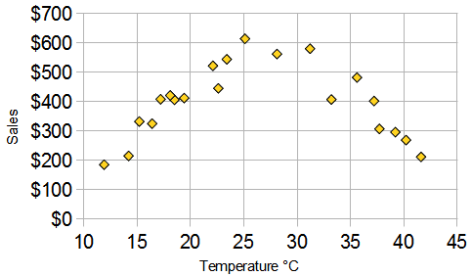
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far



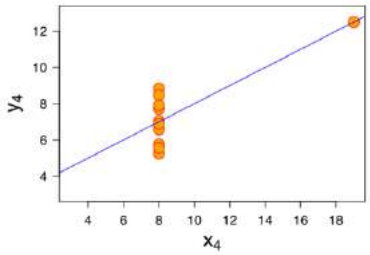
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3

Be aware of outliers and high leverage points. They can have an huge effect on the regression line.



Outlier: big $| y - \bar{y} |$

Leverage: big $| x - \bar{x} |$

Influential point: big outlier and leverage

There are statistics that quantify/describe these concepts

Let’s try simple linear regression in R…

One-way ANOVA and simple linear regression

Overview

ANOVA review and pairwise comparisons

Simple linear regression

Inference for simple linear regression

Note on grades

Midterm and final exams will be 12.5% of grade each (25% total)

Final project 10% of grade

Homework 60% of grade

Piazza activity 5% of grade

Upcoming events

Homework 6 due on Sunday October 20th

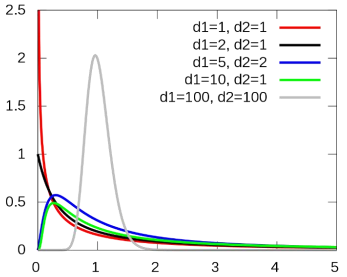
Review session during class on the 22nd

Midterm exam on the 24th

Quick review of one-way ANOVA

An Analysis of Variance (ANOVA) is a test that can be used to examine if a set of means are all the same

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- H_A: $\mu_i \neq \mu_j$ for some i, j



The statistic we use for a one-way ANOVA is the F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

One-way ANOVA table

Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{i=1}^{k} n_j(\bar{X}_j - \bar{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
Within Groups (Error)	n-k	$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$	$MSE = \frac{SSE}{n-k}$	
Total	n-1	$SST = \sum_{i=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_t)^2$		

Under H_0, these are two estimates of σ^2

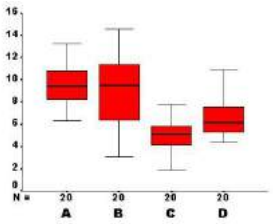
k: number of groups
n: number of data points

$$\frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Testing for equal variance between groups

Assumptions underlying a one-way ANOVA

- Data in each group come from normal distributions
- Each group has equal variance (homoskedasticity)



Can check these assumptions by:

- Visualizing data in each group
- Seeing if the ratio of $s_{max}/s_{min} < 2$

We could also run hypothesis tests to test for equal variances:

- H_0: $\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k$
- H_A: $\sigma^2_i \neq \sigma^2_j$ for some i, j
- E.g., Levene's test and Bartlet's test (Bartlet's test is sensitive to departures from normality)

Problem with the logic: if fail to reject H_0 it does not mean the σ^2's are equal, it just means we don't have enough evidence to show they are different

Non-parametric tests

There are also **non-parametric** tests which don't make assumptions about normality

The **Kruskal-Wallis** test compares several groups to see if one of the groups 'stochastically dominates' another

- Does not assume normality
- Tests whether the median for all the groups are the same
  - (if you assume groups have the same shaped and scale)
- The test is based on ranks so it is not heavily influenced by outliers

Pairwise comparisons

Usually we are interested in knowing which pairs of means differ rather than just the fact that not all the means are the same

- e.g., $\mu_G \neq \mu_{PG\text{-}13}$

There are several tests that can be used to examine which pairs of means; i.e., to test

- H_0: $\mu_i = \mu_j$
- H_A: $\mu_i \neq \mu_j$

These tests include:

- Fisher's Least Significant Difference
- Bonferroni procedure/correction
- Tukeys Honest significantly different

What is the problem with testing all pairs of means?

The problem of multiplicity

Fisher's Least Significant Difference (LSD)

1. Perform the ANOVA
2. If the ANOVA F-test is not significant, stop
3. If the ANOVA F-test is significant, then can test H_0 for a pairwise comparisons using:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$

Estimate of the SE
Uses the MSE as a pooled estimate of the σ^2

Use a t-distribution with n-k degrees of freedom

Very 'liberal' tests

- Likely to make Type I errors (lots of false rejections of H_0)
- Less likely to make Type II errors (highest chance of detecting effects)

Bonferroni correction

Controls for the ***family-wise error rate***

- i.e., $\alpha = 0.05$ for making ***any*** Type I error ***over all pairs of comparisons***

1. Choose an α-level for the family-wise error rate α
2. Decide how many comparisons you will make. Call this m.
3. Reject any hypothesis tests that have p-values less than α/m
   - Pairwise tests typically done using a t-statistic, where the MSE is used in the estimate of the SE

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$

Use a t-distribution with n-k degrees of freedom

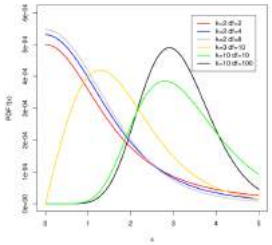
Very 'conservative' tests

- Unlikely to make Type I errors (few false rejections of H_0)
- Likely to make Type II errors (insensitive at detecting real effects)

Tukey's Honest Significantly Different Test

Controls for the family-wise error rate

$$q = \frac{\sqrt{2}(\bar{x}_i - \bar{x}_j)}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$



Where *q* comes from a ***studentized range distribution***

The test is based on the distribution of $|\bar{x}_{max} - \bar{x}_{min}|$ that would be expected under the null hypothesis that none of the pairs of means are different

- Controls for the familywise error rate but less conservative than the Bonferroni correction
- Still based on assumptions that the data in each group is normal with equal variance

Let's try the KW test and pairwise comparisons in R…

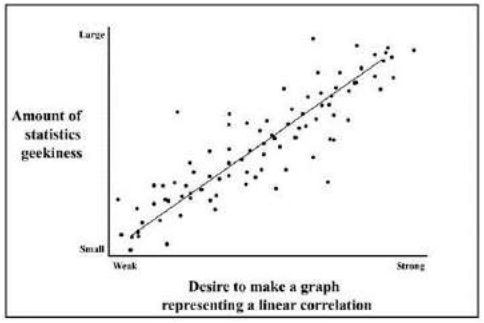
Linear regression

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

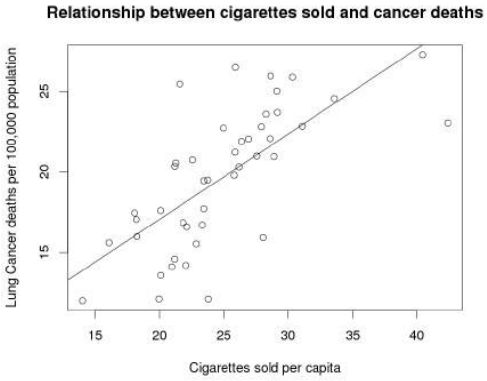
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

- In simple linear regression, we use a single variable x, to predict y



Cancer smoking regression line



$\hat{y} = b_0 + b_1 \cdot x$

`R: lm(y ~ x)`

$b_0 = 6.47$

$b_1 = 0.53$

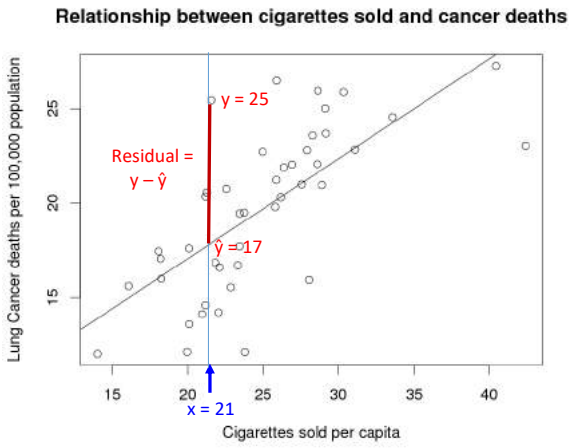
$\hat{y} = 6.47 + .53 \cdot x$

Residuals

The **residual** at a data value is the difference between the observed (y) and predicted value of the response variable

Residual = Observed – Predicted = $y - \hat{y}$

Cancer smoking residuals

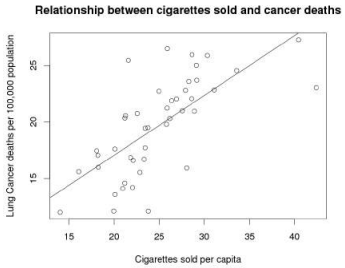


Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)
17.05	16.10	0.95
19.80	20.13	-0.33
15.98	16.12	-0.14
22.07	21.60	0.47
22.83	22.93	-0.10
24.55	24.25	0.30
27.27	27.88	-0.61
23.57	21.24	2.14

Line of 'best fit'

The **least squares line**, also called '**the line of best fit'**, is the line which minimizes the sum of squared residuals



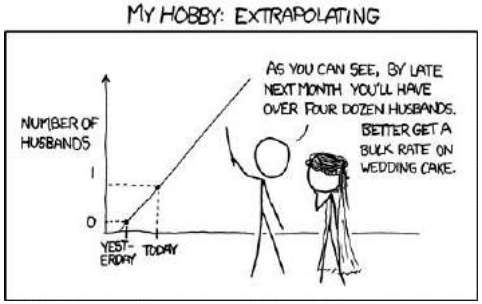
Try to find the line of best fit

Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)	Residuals² (y - ŷ)²
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

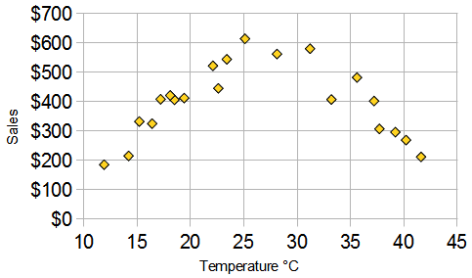
Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far



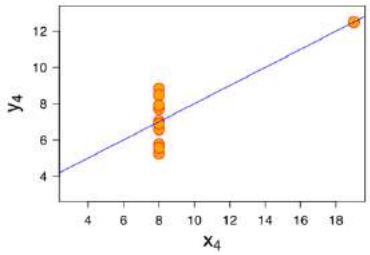
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3

Be aware of outliers and high leverage points. They can have an huge effect on the regression line.



Outlier: big $| y - \bar{y} |$

Leverage: big $| x - \bar{x} |$

Influential point: big outlier and leverage

There are statistics that quantify/describe these concepts

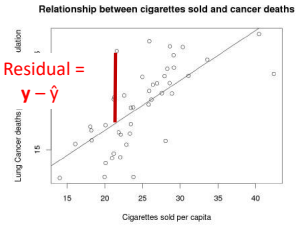
Let's try simple linear regression in R…

How do we estimate the coefficients?

Want a line that is close to our data set

- What do we mean by close?

Most commonly used measure of closeness is ***least squares* fit**, which can be calculated by minimizing the **residual sum of squares (RSS)**



$$residual = e_i$$

$$RSS$$

How do we minimize the RSS?

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x)^2$$

How do we find $\hat{\beta}_0, \hat{\beta}_1$?

Calculus and linear algebra:

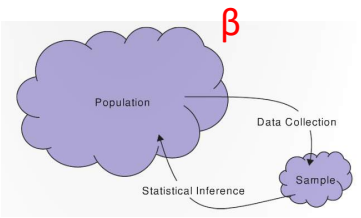
- Take the derivative, set to 0 and solve
  - This mathematical convenience is why the squared loss is so commonly used

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Inference on simple linear regression

The letter **b** is typically used to denote the slope of the sample

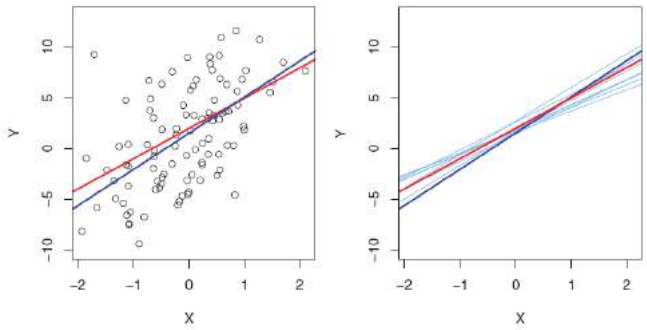
The Greek letter **β** is used to denote the slope of the population



b

Population: β

Sample estimates: b

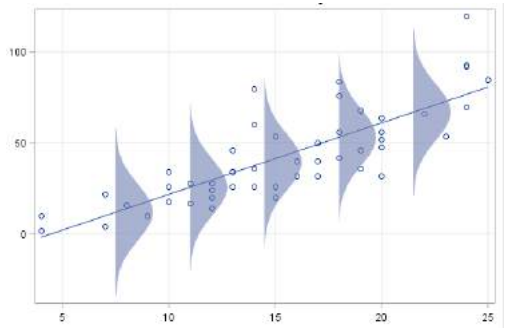


Simple linear regression underlying model

Intercept Slope } *Parameters*

$$Y \approx \beta_0 + \beta_1 x \qquad Y = \beta_0 + \beta_1 x + \epsilon$$

$$\epsilon \sim N(0, \sigma_\epsilon)$$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

Estimating σ_ε

We can also use the **residual standard error (RSE)** as an estimate standard deviation of irreducible noise $\boldsymbol{\sigma_\varepsilon}$

$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

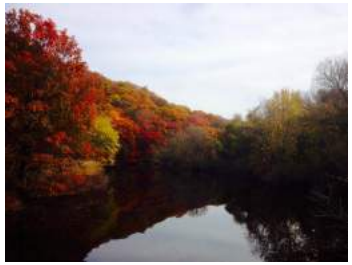
$$SE(\hat{\beta}_1)^2 = \left[\frac{1}{n} + \frac{\sigma_\epsilon^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2)}\right] \qquad SE(\hat{\beta}_0)^2 = \sigma_\epsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2)}\right]$$

Have a good fall break!

Homework 6 due on Sunday October 20th

Review session during class on the 22nd

Midterm exam on the 24th



Inference on simple linear regression and multiple regression

Overview

Data wrangling with dplyr

Review of simple linear regression

Inference for simple linear regression

Regression diagnostics

Plan for the semester (subject to change)

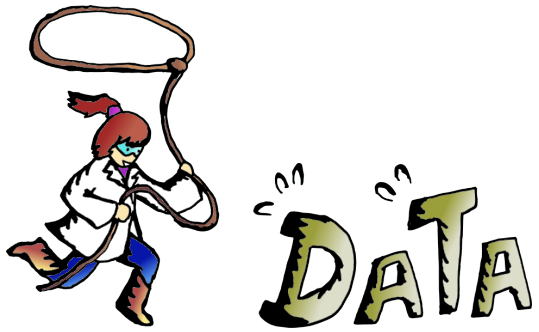
First half: topics in traditional stats analyses

Second half: Data Science approaches

What holiday is coming up?

Before we go back to regression let's eat some candy!

Data wrangling



The ‘tidyverse’

The tidyverse is set of R packages that operate ‘tidy data’

- i.e., that operate on data frames (or tibbles)

Tidy data is data where:

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

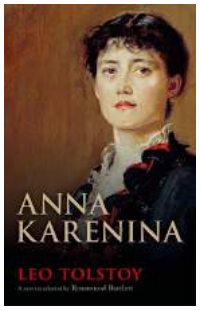
Messy data…

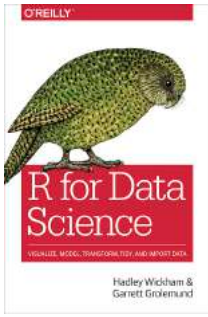
What would be an example of data that is not tidy?

Curve information - Curve quality data												
Name	Formula	Slope at	Intercept	ED-20	ED-50	ED-80	Correlatio	Forced through origo				
Standard	Calc 1: C	standarc	standarc	3792394	27752	0.2	0.5	0.8	1	No		
Plate information												
Plate	Repeat	Barcode	Measurec	Chamber	Chamber	Humidity	Humidity	Ambient	Ambient	Formula	Measurement date	
1	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	Calc 1: C	standarc	standarc 10.12.2013 10:23:33
Background information												
Plate	Label	Result	Signal	Flashes/1	Measime	MeasInfo						
1	PicoGree	0	110307	10	0	De=1st Ex=Top Em=Top Wdw=N/A						
Calculate	standarc	standards on each plate) where Label: PicoGreenFilterTop(1) channel 1										
	1	2	3	4	5	6	7	8	9	10	11	12
A	-0.0011	-0.0011	-0.001	-0.001	-0.0011	-0.0012	-0.0011	-0.0011	-0.0012	-0.0012	0.9973	1.0026
B	0.0012	0.0014	0.0013	0.0012	0.0013	0.0012	0.0014	0.0003	-0.0011	-0.0011	0.0981	0.103
C	0.0016	0.0013	0.0013	0.0011	0.0012	0.0015	0.0016	-0.0004	-0.0011	-0.0011	0.0104	0.0095
D	0.0019	0.0024	0.0018	0.0015	-0.001	-0.001	-0.001	-0.001	-0.0011	-0.0011	0.0008	0.0009
E	-0.001	-0.0011	-0.0011	-0.0011	-0.001	-0.0012	-0.0011	-0.001	-0.0009	-0.0011	-0.0001	-0.0002
F	-0.001	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.0011	-0.0009	-0.001	-0.001	-0.0003	-0.0002
G	-0.0011	-0.0011	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.001	-0.001	-0.0011	-0.0002	0.0012
H	-0.0011	-0.0012	-0.0011	-0.001	-0.0011	-0.0011	-0.0012	-0.0011	-0.0011	-0.001	-0.0003	-0.0003

Messy data…

“Happy families are all alike; every unhappy family is unhappy in its own way.”
— Leo Tolstoy



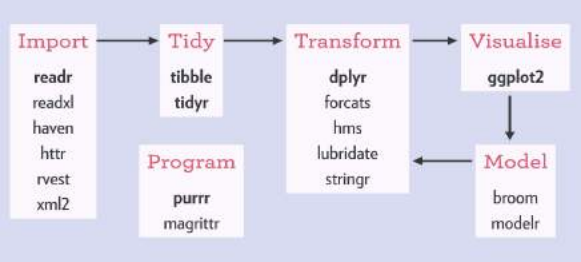


“Tidy datasets are all alike, but every messy dataset is messy in its own way.”
— Hadley Wickham

The ‘tidyverse’

The packages share a common design philosophy

- Most written by Hadley Wickham



A grammar for data wrangling

Grammar: here we mean a set of components that can be put together to achieve a goal

dplyr is a package in the tidyverse that has a set of verbs that are useful for wrangling data:

1. filter()
2. select()
3. mutate()
4. arrange()
5. group_by()
6. summarize()

All these function **take a data frame** and other arguments and **return a data**

```
> library(dplyr)   # load the dplyr package
```

Let’s look try out dplyr using faculty salary data…

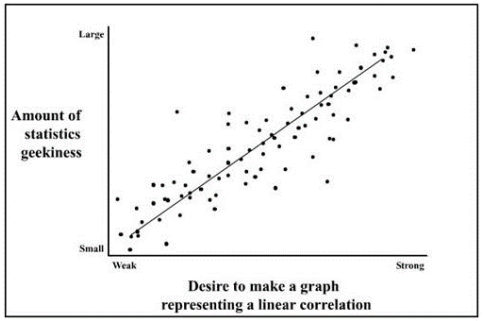
Back to linear regression

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

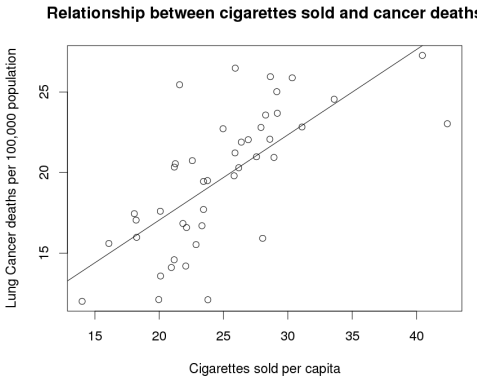
- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

- In simple linear regression, we use a single variable x, to predict y



Cancer smoking regression line



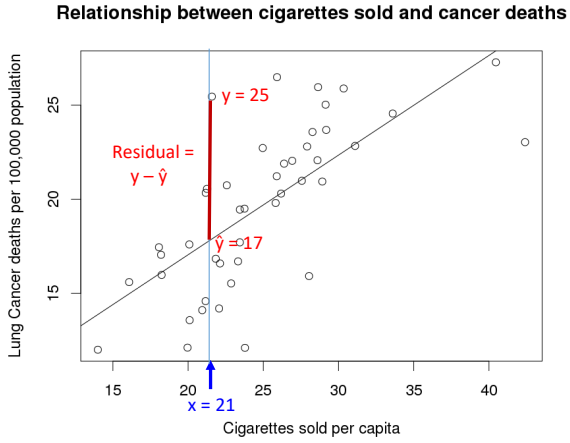
$\hat{y} = b_0 + b_1 \cdot x$

R: `lm(y ~ x)`

$b_0 = 6.47$

$b_1 = 0.53$

$\hat{y} = 6.47 + .53 \cdot x$



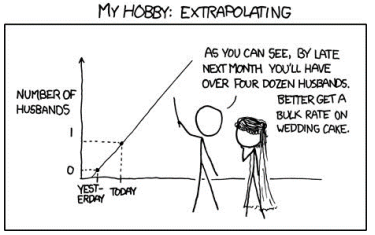
The **least squares line**, is the line which minimizes the sum of squared residuals

Residuals

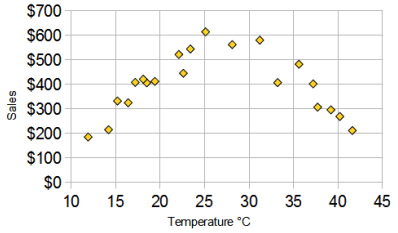
The **residual** at a data value is the difference between the observed (y) and predicted value of the response variable

$$Residual = Observed - Predicted = y - \hat{y}$$

Regression caution #1: Avoid trying to apply the regression line to predict values far from those that were used to create the line.

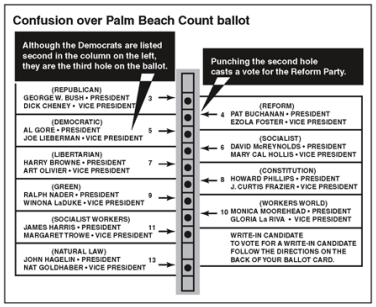


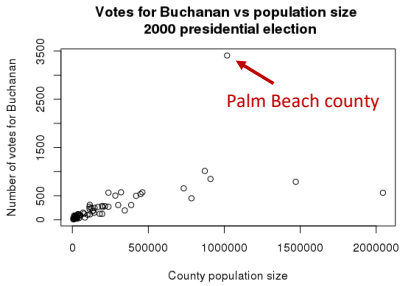
Regression caution #2: Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3: Plot the data! Regression lines are only appropriate when there is a linear trend in the data.

The butterfly ballot

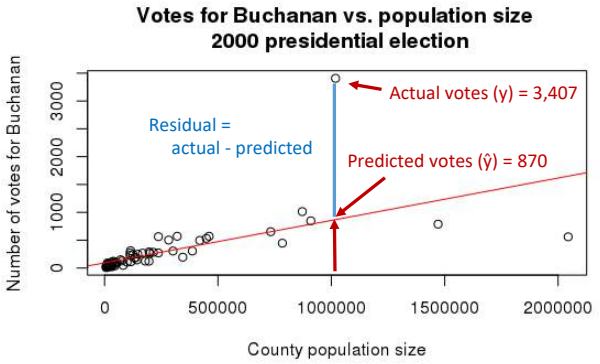




Q: What do we do when we have outliers?

A: Investigate!

Regression analysis

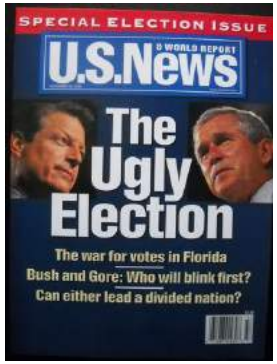


What is the residual for Palm Beach county?

Actual - predicted
= y - ŷ
= 3,407 - 870
= 2,537 votes

i.e., Gore likely should have had ~2,500 more votes

Side note: what was the outcome of the 2000 presidential election?

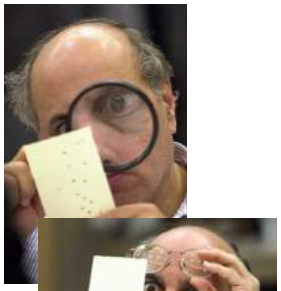


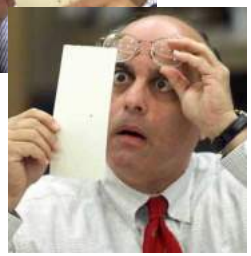
<u>Official final Florida vote count:</u>
Total votes Bush = 2,912,790
Total votes Gore = 2,912,253

Office vote count difference:
= 2,912,790 - 2,912,253
= 537

Residual votes: 2,537

Conclusions?

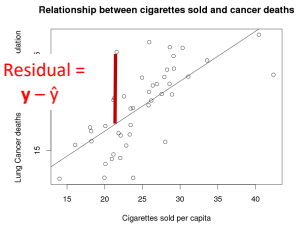




How do we estimate the coefficients?

As mentioned before, we minimize the **sum of squared errors (SSE or SSError)** to calculate the least squares regression line.

- The residual sum of squares is also called the **residual sum of squares (RSS)**



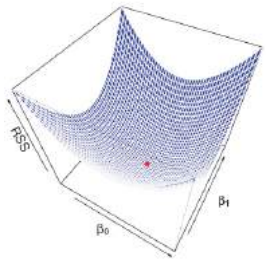
$$residual = e_i = y_i - \hat{y}_i$$

$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{f}(x))^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x)^2$$

How do we minimize the SSError?

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x)^2$$



How do we find $\hat{\beta}_0, \hat{\beta}_1$?

Calculus and linear algebra:

- Take the derivative, set to 0 and solve
  - This mathematical convenience is why the squared loss is so commonly used

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad r = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \hat{\beta}_1 \cdot \frac{s_x}{s_y}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

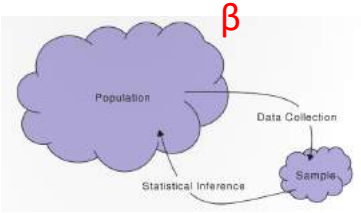
Side note: does this equation look somewhat familiar?

Inference on simple linear regression

The letter **b** or $\hat{\beta}$ is typically used to denote the slope of the sample

The Greek letter **β** is used to denote the slope of the population

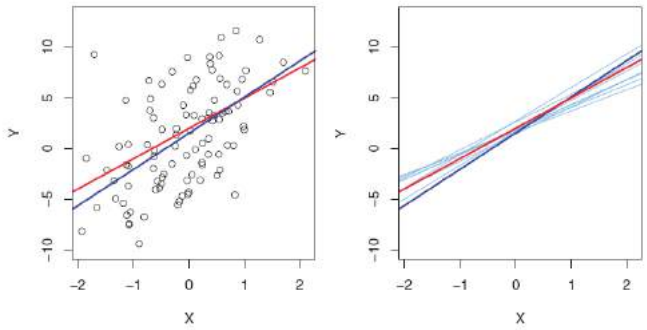
β



b $\hat{\beta}$

Population: β

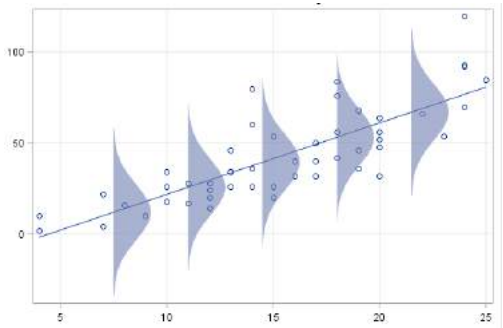
Sample estimates: b $\hat{\beta}$



Simple linear regression underlying model

Intercept Slope } *Parameters*

$$Y \approx \beta_0 + \beta_1 x \qquad Y = \beta_0 + \beta_1 x + \epsilon$$



$$\epsilon \sim N(0, \sigma_\epsilon)$$

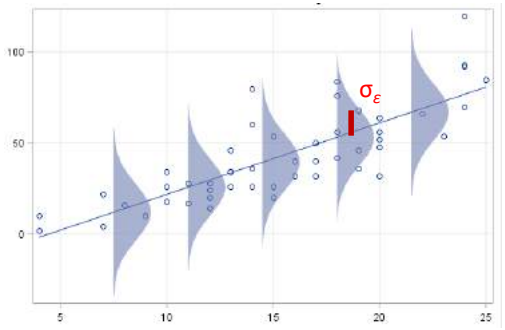
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

Estimating σ_ε

We can also use the **standard deviation of errors** as an estimate standard deviation of irreducible noise $\boldsymbol{\sigma_\epsilon}$

- This is also called the **residual standard error (RSE)**



$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2}SSE}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

One type of hypothesis test we can run is based on a t-statistic: $t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$

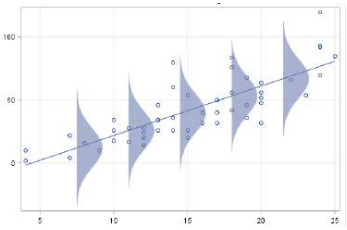
- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

$$SE_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}} \qquad SE_{\hat{\beta}_0} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2)}}$$

Inference using parametric methods

When using parametric methods, we usually make the following assumptions:

- **Normality**: residuals are normally distributed around the predicted value ŷ
- **Homoscedasticity**: constant variance over the whole range of x values
- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points



These assumptions are usually checked after the models are fit using 'regression diagnostic' plots.

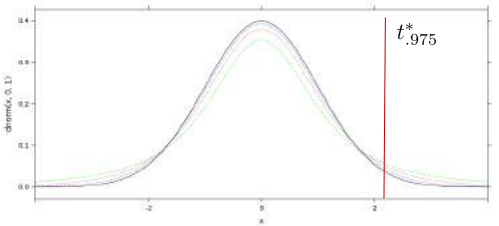
Confidence intervals for regression coefficients

For the slope coefficient , the confidence interval is: $\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$

Where: $SE_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$

t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence

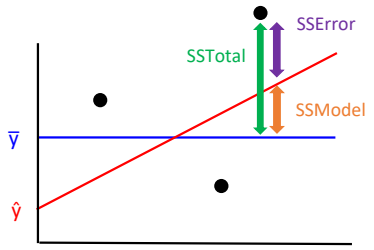
N(0, 1)
df = 2
df = 5
df = 15



Hypothesis test based on ANOVA for regression

For simple linear regression, we can also run a hypothesis test on the regression coefficients based on an analysis of variance (ANOVA)

- For simple linear regression, this gives exactly the same results as running a t-test. $F = t^2$



The ANOVA decomposes the variance as:

- SSTotal = SSModel + SSError

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Added and subtracted ŷ

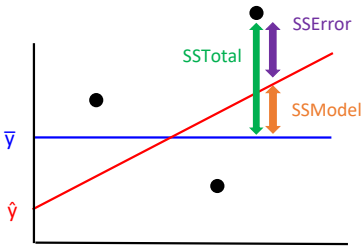
$$(y - \bar{y})^2 = (\hat{y} - \bar{y})^2 + (y - \hat{y})^2 + 2 \cdot (y - \hat{y})(\hat{y} - \bar{y})$$

This equal 0 (proof via algebra)

Hypothesis test based on ANOVA for regression

The **percentage of the total variability explained by the model** is given by

$$r^2 = \frac{SSModel}{SSTotal} = 1 - \frac{SSError}{SSTotal}$$



The ANOVA decomposes the variance as:

- SSTotal = SSModel + SSError

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Added and subtracted ŷ

$$(y - \bar{y})^2 = (\hat{y} - \bar{y})^2 + (y - \hat{y})^2 + 2 \cdot (y - \hat{y})(\hat{y} - \bar{y})$$

This equal 0 (proof via algebra)

Hypothesis test based on ANOVA for regression

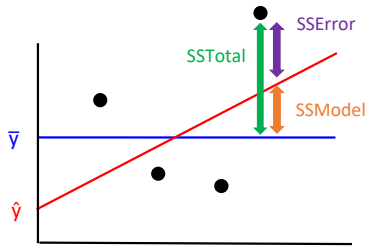
$$F = \frac{SSModel/df_{model}}{SSError/df_{error}}$$

$df_{model} = 1$

$df_{error} = n - 2$

If the null hypothesis is true that $\beta_0 = 0$:

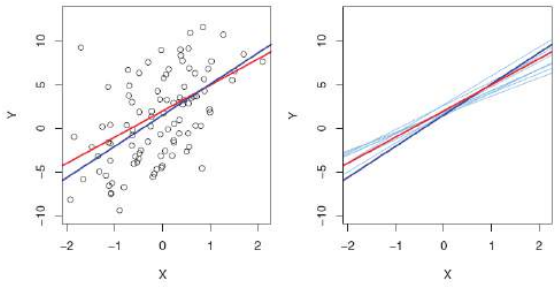
- Both the numerator and denominator are estimates of σ^2
- F comes from an F-distribution with df_{model}, df_{error} degrees of freedom



Resampling methods for inference in regression

We can also use resampling methods to estimate run hypothesis tests and create confidence intervals for the regression coefficients

- Bootstrap
- Permutation test



Let’s look at inference for simple linear regression in R

Using faculty salaries…